WP. 17
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)

Topic (iii): Editing microdata for release

EDITING AND IMPUTATION OF LINKED MICRO FILES TO BE USED IN STATISTICS, RESEARCH PROJECTS AND ADMINISTRATIVE PROCEDURES

Invited Paper

Prepared by Sindre Børke (sbe@ssb.no) and Svein Gåsemyr (sga@ssb.no), Statistics Norway

I. INTRODUCTION

1. The infrastructure that is in place in order to produce official statistics based on administrative data also forms the basis for production high quality micro data for research purposes. The quality controls needed for official statistics are not necessarily the same controls that are needed for research purposes. Research projects are often concerned with analyzing *small groups* and applying data to *longitudinal studies*.

2. For more than half a century Norway and the other Nordic countries have developed an advanced infrastructure to promote electronic reporting from enterprises and households to government sector. The infrastructure consists of administrative base registers for person, enterprise and building/dwelling. The base registers assign unique and official ID numbers to be used in government and private sector. At the moment there is a strong development in common portals for Electronic Data Interchange, "Altinn" for businesses and household. A common security infrastructure (PKI) for eGovernment is under development. The aim of this infrastructure is to reduce response burden and ensure efficient data reporting.

3. The Nordic NSIs have drafted a report on register-based statistics, February 2006. The final report is to be published by UNECE, [1].

4. A report initiated by the Research Council of Norway on the need to improve the infrastructure within social scientific research, particular with regard to access to register data for research purposes, was published in 2003, [2]. An infrastructure should be developed in Statistics Norway that safeguards effective procedures for ordering and delivering data and for feedback on experiences in the use of data. Access to information on what type of data is actually available should be made easier. Work should be intensified in *consistency controls* and *mapping of errors* sources in popular data systems. Access to data should in principle be free, but Statistics Norway should be able to charge research environments if data delivery entails a large degree of linked files beyond a defined set of standard registers. The number of different data sources that could be linked is about 100 administrative data systems and 100 statistical surveys.

5. The proposals to improve the infrastructure for creating micro files for statistics and research projects are followed up by a large project supported by the Research Council and Statistics Norway. The

project started in 2005 and will be finished in 2006. The paper presents results from this project and concentrates on detecting and correcting for inconsistencies when linking different sources at micro level.

II. EDITING, IMPUTATION AND ELECTRONIC REPORTING

6. So far the administrative data systems for the units of person and household have been the most important sources for statistics. At the moment there is a strong development in administrative data for the units of enterprise and establishment. Administrative data on enterprise accounts is now available for almost all enterprises also for farmers and family companies. This trend is combined with more and more use of electronic reporting by enterprises and more coordinated reporting from enterprises to government agencies incl. statistical surveys. Computer assisted methods for editing and imputation would be based on combined use of multimode business surveys, (paper forms, electronic questionnaires and direct reporting from internal administrative systems), surveys of previous periods and administrative data.

A. Electronic data collection

7. With electronic questionnaires and electronic data interchange (from the enterprise's data systems) both the respondent and the collector will expect some checking and editing to be integrated at response time. If the wanted information for a business survey not is available in administrative data system of the enterprise there would be problems and the response burden might increase. Statistics Norway has started a project on electronic reporting from large enterprises to increase our competence on enterprise internal data systems.

8. The control checks within the procedure of electronic reporting, will, in principle, be the same as used on the input from paper questionnaires. The difference being that this editing is placed nearer to the source, at the respondent.

9. Usually, the data sent after editing from the respondent is the only data registered. Using client side Paradata (process data), we are able to observe the communication between the questionnaire and the respondent. This will add information on how the questionnaire is working, but also how the respondents change their answers when a notice is given. This information might be used in editing data from the paper-questionnaires in the same survey. So far, no attempts are done to explore this. A warning is registered, observing that there seems to be more errors identified in web-questionnaires than in paper-questionnaires in the same survey, [5].

B. Editing using data from different sources

10. When constructing controls to be executed at response time, you might include information from different sources, known to the collector. So far, these controls have been limited due to several reasons. The complexity may grow very high, the response burden too. Even response time from the computer might pose a limitation. Often the survey is held before information from an administrative source is available.

11. Once the data is captured, there will be an opening for more complex editing and imputations, based on information from several sources. One possible source, the Paradata from web-questionnaires is mentioned above.

12. A database on enterprise accounts is in operation. It is proposed to include all available data of the units of enterprise and establishment. This will be a fundament also for editing and imputation purposes for each survey's data and the survey data can be added to the dataset of the base.

C. Dynamic revision system

13. As a step in this direction Statistics Norway develops a dynamic editing system (Dynarev). In this system there are possibilities to check for inconsistencies in each unit, compared with other variables

in the same questionnaire, but also in already known information from earlier surveys or register information. The plan is to develop Dynarev to utilise all available data for enenterprise and establishment.

14. Partly in Dynarev and partly in general, the editing procedures should include different statistical controls, focusing on critical variables and units in the survey. The identification of unusual changes from earlier data on unit level is a part of this. Editing using graphs is included in the concept. Further, the Dynarev facilitates controls on aggregated data, identifying variables that seems to have unexpected changes, following up by inspections on underlying data.

15. In Dynarev, as in most systems for editing and imputation, there will be registered some process variables. Both statistics on the editing process and identifications of changes being done are important. Statistics Norway is trying to improve the reporting of such information through a standard set of quality indicators for the data collecting process.

16. Dynarev is developed for, and until now used in a few surveys on prize indexes. Statistics Norway will be able to present this system in a later Work Shops.

III. DATA SOURCES OF INTEREST FOR RESEARCH

17. Administrative base registers and data systems are usually continuous updated and can be organized as longitudinal databases. Very often research projects are based on longitudinal data and this paper discusses the most important longitudinal databases for research purposes.

A. Base registers

18. Administrative and statistical base registers on person, establishment and building/dwelling-/address are organized as longitudinal databases. A common data model is developed for the 3 statistical base registers in Statistics Norway.

B. Administrative data systems on person

19. The most important databases for research should be:

- Current and completed educational programs and educational attainment of person
- Job, spell of unemployment, in labour market measure
- Income account of persons
- Data systems of Social Security, pensions and transfers
- Annual file for Population and Housing Census (under development)
- There are data sources to create an integrated database on health indicators

C. Administrative data systems on enterprises

20. So far most register-based statistics are statistics on person and family. At the moment we see a strong development in use of administrative data systems on enterprise and establishment as a source for statistics. Enterprise accounts are now available for most enterprises. A database on enterprise accounts is under development. It is proposed to include other administrative sources in this database such as VAT, wage sums and employment.

21. A good infrastructure to link 3 base registers and 7 data systems could be to apply the same data model for all longitudinal databases and create a menu in such a way that the researcher should specify the needed variables and period of time. The aim to develop an infrastructure in Statistics Norway that it should be easy and cheap to create the specified micro file.

D. Statistical surveys

22. Usually micro files of statistical surveys are linked to the actual base registers and other administrative data systems.

E. To follow statistical units through time

23. One advantage of administrative data is that a unit could be followed through time. Some units are easy to follow, e.g. a person. One of the objectives of the Business Register is to follow the unit of establishment over time. The CPR covers the period after 1960, the initial situation is the Census 1960. Generation files are now created for research project, i.e. units of brothers and sisters and cousins and nephews are followed for a period. The initial file for the database on fulfilled educational programs and educational attainment is the information of educational attainment in the Census 1970. Some information on annual income is from 1967, more detailed information starts in 1992. The first register-based micro file for Population Census is from 2001.

IV. THE INFRASTRUCTURE IN STATISTICS NORWAY FOR CREATING LINKED MICRO FILES FOR RESEARCH PROJECTS

24. Empirical research within economy and other social studies has to an increasingly greater extent gravitated towards micro data on persons and enterprises. Such data can originate from sample surveys or be retrieved from base register and other administrative data systems. Administrative data have the obvious advantage that they provide far more data material, and often include the entire relevant population.

25. In resent years it has become steadily more common for data for scientific analysis to be generated via linking various registers. This opens up a wealth of analysis potential. First, it provides the opportunity to simultaneously observe many more variables for the same unit. Second, links to registers for different years make it possible to follow units over a period of time, whereby different courses of progress can be studied in rich and extensive data sets. Furthermore, linking registers for different types of units, e.g. persons and enterprises, makes it possible to study the interaction between these.

26. Linking registers in order to obtain a broader set of variables is one of the most important benefits of using register data in research, and opens up a number of new possibilities for analysis. Use of linked files means more work in order to *control the consistency* between the registers. This can often require considerable effort, not least because the scope of different registers can vary, the use of ID numbers can be inconsistent between different registers, and the registers can have different reference dates. In many applications the benefit of using administrative data in research and analysis is related to being able to follow persons and enterprises over a period of time. This requires resources in the form of establishing databases with long time series of micro data. This is work in which the benefits are only fully reaped in the long term. The product has the hallmark of being a public good. The complexity of this work indicates that that a large part of the *basic work linked to establishing register-based databases should be centralized.* Documentation of data sources is important. In order to be able to use the full research potential in administrative data it is necessary to know what type of data sources exist and what type of variables they contain. There are clear adventages to such documentation work being centralized, and for the documentation being available across the boar.

27. The report of Research Council [2], defines what should characterize a good infrastructure linked to the use of register data. This is detailed in six points. Point two refers to standardized identification of units, the possibility of linking units and quality controls and consistency checks of linked files.

V. EDITING AND IMPUTATION OF BASE REGISTERS, OTHER ADMINISTRATIVE SOURCES AND SYSTEMS OF LINKED FILES

28. As there are more than 100 administrative data systems and about 100 statistical surveys that can be linked, the infrastructure for linking is of importance. When linking data systems, researchers are faced with a range of possible combinations of values on variables that can seem inconsistent. Inconsistency could be a result of regulations that apply, inaccurate indication of reference period, direct measurement errors and registration errors. An example of inconsistency, a person is registered as full time employee in the data system of Social Security but without labour income in the data system of Tax Agency. Adjustment to ensure consistent data is needed where the data is to be used for research purposes. This requires data knowledge with regard to how the data is registered and how data can be combined.

29. The quality control and consistency checks that are carried out for official statistics are often insufficient in relation to the researchers' needs. In addition to consistency checks for standard cross-sectional data controls of micro data across subject areas are needed. Few attempts have been made to co-ordinate the data needs of the various research environments.

30. The inquiry team of the Research Council believes that resources should be set aside for extended consistency controls, quality checks and the co-ordination of areas in which the research has special needs. The work should particularly be concentrated on the set with base registers. This work should be managed by and carried out by Statistics Norway in close communication with central research environments. Together with the research environments, it should be determined where the need is greatest in relation to consistency, quality and co-ordination, and projects should be initiated that tackle the most pressing problems.

31. One example of such work is the “Dynamic data person-establishment” project, which was supported by The Research Council of Norway through the “Recruitment and labour market” programme. Experiences from this project are useful in relation to how such work can be carried out. The project was managed and executed by Statistics Norway, but had a reference/steering group with representatives from the research environments that had most knowledge in the use of the relevant data sources.

32. This section concentrates on longitudinal databases that are of importance for research projects. The problems are related to cases of inconsistency within each of the longitudinal databases and when a specified micro file is to be based on linking of several databases

A. Longitudinal data

33. Administrative data systems are usually continuous updated and can be organized as longitudinal data, i.e. the unit is followed over time. Longitudinal data means that the period of which a specific value of a variable is valid is specified. Consistency controls for longitudinal data are more complicated than for data for a specified reference period like a date or a year. Events have to be specified in correct order for each unit.

B. Base registers

34. A project in Statistics Norway to improve the infrastructure of the statistical base registers started in 2004. The aim is to improve the services of 3 base registers. A common data model is developed. This model could be utilized also for the other longitudinal databases to easy the linkage of about 10 databases. An important control of all kind of longitudinal databases is to check that there is a correct chronology of the registered events.

35. Efficient linkage of the 3 base registers at unit level is an important quality. Official addresses is registered in GAB and is used in the CPR and BR. The unit of job is identified both with the PIN of the

employed and the BIN of the work place and has a key role in development of an integrated statistical system. For this reason Statistics Sweden consider the job file as the fourth base register.

C. Central Population Register (CPR)

36. Reports to the population registration are based on legal events, (birth, marriage, death etc.). This reporting is of good quality and fast. The family reports on migration to a new address, usually there are some delays in the reporting. Fore some cases the delay is several weeks. The result is inconsistency between data on residence and workplace. Some families do not report emigration. The result is that the family is registered as resident in the CPR, but not registered in any other administrative data systems e.g. on activities or income.

D. Dwelling Register

37. The Housing Census 2001 contributed to the establishment of the initial micro file for an administrative base register on dwelling. The ID number of dwelling is registered as a variable in the CPR. This variable is the source to develop register-based household statistics.
Unfortunately about 5% of the families are missing ID number of dwelling. Statistics Norway started publishing register-based statistics on couples, family and household this year. The statistics are based on two systems of imputation. The missing ID number of dwelling is imputed based on information of address. As there is no reporting about cohabiting couples this unit is created by imputation on the base of information of household composition. The imputation is based on statistical models that ensure correct statistics. If the unit of imputed couples is to be used in microanalysis it should be useful to know how many of the imputed couples are real couples.

E. Business Registers

38. The quality of the BR has to be very good when used in projects to study the dynamic relations between employed persons and establishment. The importance of quality refers to aspects such as information on the period of which an establishment is economic active, that the BR covers all sectors and industries, and the definition of the unit of establishment through time. Statistics Norway publishes demographic statistics of the unit of establishment based on the BR. A prerequisite for this statistics is a good quality of the BR.

F. Database on enterprise accounts

39. Up to now the register-based statistics is most developed for statistics on persons. In the last years more and more administrative sources for the units of enterprises and establishments are available. One important component of this development is that it is very close to have annual enterprise accounts for all enterprises also for small family companies and farmers. It is proposed to develop the database for enterprise accounts to include all existing sources, VAT, wage sum, employment, and statistical surveys for the units of enterprise and establishment into one database. These sources have to be in consistency and the database would be very useful inputs for the procedures of editing and imputation.

G. Database on job, spell of unemployment, and in labour market measure

40. Method for editing and imputation of a linked job file was the main topic of a Norwegian paper to the editing WS in Ottawa 2005, [4]. The Ottawa paper presents results from the work on the Census micro file and the “Dynamic data person-establishment” project.

41. Micro files on the dynamic relations between employed persons and workplaces have been created for several projects. These relations means that all variable for an employed can be linked to the unit of establishment staff and variables of an establishment can be linked to micro files on persons.

42. When two sources are linked at unit level we always find some cases of inconsistency. And in a case of inconsistence it could be difficult to see if the reason is difference in definition, different reference period or an error in one of the source.

43. An example of inconsistency of a linked file of the data system of employee jobs and the CPR: Some of the employees are dead according to the CPR or emigrated to a foreign country.

44. Status in labour market is classified by a set of rules. The rules are based on thoroughly knowledge about the quality of each source.

Table 1. Employed persons by data source for status in the labour market 4th quarter

Source	2003	2004
Employed persons	2 260 000	2 274 000
Employee jobs both in social security and Tax reg.	1 897 700	1 911 200
Employee jobs only in social security register	24 500	17 200
Employee jobs only in Tax register	169 400	181 700
Self-employed primary industry	52 400	46 200
Self-employed other industries	109 600	111 800
Conscripts	6 300	5 900

H. Annual micro file for population and housing census

45. An annual micro file for population and Housing Census is in operation. Some important census variables are based on more than one of the databases described in this section. The census file includes units of *time use activities* such as: paid work, spell of unemployment, period in a labour market measure, period in current education units of *sources of livelihood* income such as: labour income, sickness and unemployment benefit, disabled and old age pension, education grant, student loan. Main and secondary time use activities and main and secondary sources of livelihood should be classified.

46. The subject matter unit of a database has the responsibility for editing and imputation. When several databases are linked at micro level the unit of Census is responsible for the editing of cases on inconsistency between databases. A set of rules decides which of the database should be the most reliable. Several conditions are involved in the process.

I. Time use activities and sources of livelihood organized as longitudinal data

47. Reference periods of the annual census file are restricted to i) the calendar year and
ii) the first week of November. In micro file for research projects the Census variables should be organized as longitudinal data. Pensions in Norway usually refer to a month, i.e. the status starts the 1st of a month and ends the last date of the reference month. Integration at micro level of longitudinal databases might be a difficult task. Some researchers might want to take care of editing and imputation themselves, but it seems that the research environment would like Statistics Norway to develop a sufficient infrastructure for integrating longitudinal databases.

J. Micro files on generations

48. The CPR registers demographic events from the period after 1 November 1960. By use of the variables ID of mother and father units such as brothers and sisters, grand children can be followed through time. The Norwegian CPR dos not register biological parents, but this is the case in the CPR of

other Nordic countries. Several research projects are based on micro files where generations are followed for a long period.

VI. EDITING AND IMPUTATION OF LINKED FILES FOR ADMINISTRATIVE DATA SYSTEMS

49. This year the Norwegian government agencies of Social Security, Employment Service and the municipal offices of Social Assistance are integrated into one agency. The new agency is to be developed within a period of 5 years. The Norwegian labour force is about 2.4 mill. The number of non old age pensioners, not in employment and with some kind of transfer as main source of livelihood is about 700 000. The aim of integrating the agencies responsible for income security, labour market and social service is to increase employment and reduce the number of (i) unemployed, (ii) employed on sickness absenteeism and (iii) not pensioner and not in the labour force.

50. The development of a more efficient system to integrate persons with weak health condition and social problems as permanent members of the labour force needs a very advanced system of integrated databases. The service of the new agency is to be based on electronic casework.

51. In the first step the casework is to be based on the existing administrative data systems of the former agencies and the use of a menu that can select information from all current updated existing data systems. In the medium term the plan is to develop one common database to provide all needed information for any type of client. This database will contain units like job, spell of unemployment, period in labour market measures (job training or education), absent from work (sickness, lay of for delivery, and other), medical rehabilitation, labour market rehabilitation, health indicators, application for disabled pension, rejection of disabled pension, temporary disabled pension, and disabled pension. For each of these units or statuses there is an income source, (labour income or transfer). The data have to be organized as a longitudinal database for a rather long period. For some cases the needed information should cover the life course of a person. There have to be a current updating of this database.

52. This database is very close to the database researchers in labour market and living condition are interested in. Concerning methods of editing and imputation there is an important difference between an administrative database and a database for statistics and research. In the database for statistics and research corrections to ensure consistency within a single source and within a linked file usually is corrected by a set of rules. Inconsistency within an administrative database has to be corrected by some kind of documentation of what should be the correct data source or value of a variable. The client could also tell what should be the correct information.

VII. CONFIDENTIALITY AND PRIVACY

53. Register data is often confidential or sensitive. This put restrictions on the access to and use of the data. Safeguarding access to micro files whilst complying with privacy protection and confidentiality provisions requires significant resources and good procedures by those supplying the data and those using the data for research.

54. In Norway micro files are normally made available whereby they are delivered to the individual researcher/research institute after the data have been anonymised or de-identified and legal requirements have been met. In general, the following classification of personal data applies in registers.

(i) Identifiable - the person is identified by variables such as ID number, name and address.
(ii) De-identified - ID number, name and address are removed, but it is still possible to identify individuals via other variables when linking to other sources.
(iii) Anonymised - in practice, it is not possible to link data in the micro file to individuals.

55. All variables of the micro file have to be taken into consideration. Anonymised data is not regarded as personal data. The Data Inspectorate of Norway does not put direct restriction on the delivery for research projects. Personal data can be delivered where the following conditions are met:

- Concession for the Data Inspectorate is needed
- The data have been de-indentified to the degree that the purpose allows
- If the data are subject to confidentiality provisions outside the Statistics Act, the researcher must obtain dispensation from the duty of confidentiality.

56. One crucial objective is to give researchers and students the easiest access possible to as much of Statistics Norway's data as possible within the regulations that apply for the protection of privacy and confidentiality. Statistics Norway and Norwegian Social Science Data Services (NSD) have a far-reaching collaboration, and NSD is an important organ in the forwarding of Statistics Norway's data. Collaboration between the two institutions is an important premise for this. The agreement regulates the delivery of anonymised data from NSD for research. In a number of cases, however, problems and research methods dictate that researchers receive de-identified data. As a main rule, the researchers in these cases are served directly by Statistics Norway. In some cases, separate agreements are entered into between Statistics Norway and NSD for NSD to provide researchers with de-identified data.

57. Stringent requirements are set for confidentiality and integrity when using register data, and the more identifiable and sensitive the data are, the more stringent the requirements are. Provisions have been established, however, in many parts of the research sector, which safeguard such considerations with regard to ethics, regulations and formal requirements, and information and guidance on such issues. Examples of such provisions are the ethics committees, The Ombudsman for Privacy in Research and the Council of Confidentiality. Although the requirements mean extra resource use for research, the provisions are crucial to safeguarding important privacy protection interests and to maintaining a good reputation and confidence in research in general.

References

[1] Register-based statistics in the Nordic countries - documentation of best practices, draft, February 2006, to be published by UNECE.

[2] Infrastructure in social science - Access to register data for research purposes, Report for The Research Council of Norway by an inquiry team, Oslo 2003.

[3] Svein Gåsemyr (2005): Record linking of base registers and other administrative sources - problems and methods. Paper to Siena Group meeting Helsinki, February 2005.

[4] Svein Gåsemyr, Editing and imputation for the creation of a linked micro file from base registers and other administrative data. Paper to the WS on editing, Ottawa 2005.

[5] Gustav Haraldsen et. al. Paradata indications of problems in Web surveys. Presentation at European Conference on Quality in Survey Statistics, Cardiff, April 2006.
