WP. 8
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (i): Editing administrative data and combined data sources

EDITING AND IMPUTATION FOR THE CREATION OF A LINKED MICRO FILE FROM BASE REGISTERS AND OTHER ADMINISTRATIVE DATA

Supporting Paper

Submitted by Statistics Norway[1]

I. INTRODUCTION

1. When using administrative data for statistical purposes editing (checks and corrections) has to be based on computerized procedures. The volume of data (records x variables), of administrative sources is simply too large to base the editing on manual methods alone. Usually the micro file of a statistical survey is linked to base registers and other administrative sources. The advantage of integrating surveys and administrative sources is to utilize the strength of each source. This paper concentrates on editing and imputation of linked files from base registers and other administrative sources.

2. The procedures of editing and imputation are integrated parts of data collection and processing. Section II of the paper describes the infrastructure that is developed to promote electronic collection and processing of data and reduced response burden for enterprises and households. Section III presents problems and methods in editing and imputation for creating a system of linked files by describing problems and methods in creating the job file. The job file is an integration of the following administrative and statistical units: employee jobs, self-employed jobs, spells of unemployment, and periods in a labour market measure. The integrated job file has a key role in developing integrated and coherent social statistics and in integration of economic and social statistics. The job file was developed for the Norwegian register-based Population Census 2001. The job file might be the most difficult component to develop within a system of register-based statistics. Other topics discussed in this paper are; extended job file in section IV, measuring quality of linked files in section V, and linked files for administrative purposes in section VI.

II. INFRASTRUCTURE TO PROMOTE ELECTRONIC COLLECTION AND PROCESSING OF DATA

3. Since the start of the 1960s in the Nordic countries, the development and use of computerized administrative data systems in private sector and government agencies has been based on the same infrastructure and strategy of data collection:

- Operation of computerized central administrative base registers of the total population for the most important units, (i) person, (ii) business and (iii) land property/dwelling;

[1] Prepared by Svein Gåsemyr (sga@ssb.no).

- Assignment of an official and unique ID number for the main units, PIN (Person Identification Number) BIN (Business) and DIN (Dwelling) and use of the unique and official ID numbers in administrative data systems of government agencies and the private sector and in statistical surveys;

- Infrastructure under development for a few years is operation of a common portal for EDI reporting (Electronic Data Interchange), methods for designing electronic questionnaires and common meta information systems.

4. The task of the base registers is to identify target populations. The tasks of the ID numbers are to ensure reliable identification of a unit and efficient linkage across sources at unit level and to follow a unit over time.

A. Uses of a common portal and electronic questionnaire and how these measures affect methods for editing and processing

5. Since 1996 Statistics Norway has put considerable efforts into the area of electronic data collection. Three systems have been implemented, Kostra for the municipalities, Idun for web questionnaires in business surveys and Altinn as a common portal for data reporting to the government sector. Household surveys conducted by Statistics Norway are based on CATI.

6. Electronic data collection implies that respondents are doing a large part of editing. Furthermore, the respondents have become users of a NSI system, where you introduce the need for traditional service support, and finally, when the respondents press the send button, they expect an immediate response indicating whether everything has been accepted or not. Therefore, electronic data collection will have a great impact on the internal workflow and responsibilities concerning the data collection process. These impacts will occur immediately when you provide electronic data reporting as an alternative for most of your paper surveys.

7. Mixed mode data collection will emphasize the need for a central reception of all incoming responses. The editing defined and carried out in the electronic questionnaire could be reused both on file extracts and on the paper versions of the same questionnaire during scanning and verify operations. While the respondents can decide how to react on a warning made by a programmed rule, the scanning and verify operator will not always be able to take the same decision. This introduces the need for 1) training of staff and 2) automated edits and tracking flags. The objective however, should be to hand over to the subject matter division a micro-file with as few differences as possible caused by different data collection methods. This would allow the subject matter divisions to concentrate more on data analysis and editing from a macro perspective.

8. When reporting electronically, respondents need immediate feedback when they deliver their information. This introduces the need for common procedures to verify that they have completed their reporting duty. In most cases, central staff and procedures will be sufficient. Data reported by different methods should be handled equally.

B. Main data sources for official Norwegian statistics

9. Statistics Norway (SSB) uses administrative data as a source for statistics whenever it is possible. The policy of the Norwegian government is that an enterprise should report a variable to government agencies only once. Of 200 official statistics in SSB about 100 are based on administrative data. The volume of the data (records * variables) is much bigger for administrative sources than for statistical surveys.

Administrative base registers: Central population Register (CPR). Legal Unit Register (LUR). Property and Dwelling Register (GAB).

Administrative data for both social and economic statistics: Income and wealth, labour market data.

Sources based on a combination of administrative data and surveys: data for wage statistics, Kostra (system for EDI collection from municipalities), and database for completed educational programs.

Administrative data systems to be used in social statistics: Social security, social care, health, culture and crime.

Important administrative data systems used for economic statistics: foreign trade, VAT, enterprise accounts (all enterprises are covered from 2005).

Household surveys: Labour Force Survey, Household Budget Survey, Coordinated surveys on living conditions, Survey on culture and media, Time use survey (every 10 years).

Business surveys for short-term statistics: Production and turnover of key industries, Retail trade, CPI, Investment of key industries, Orders of key industries, Use of Internet by households and enterprises.

Business structural statistics surveys: Annual structural survey of most industries.

III. METHODS FOR EDITING AND IMPUTATION OF A LINKED JOB FILE

10. Usually some editing is executed by the agency that is responsible for an administrative data system and there are some controls and editing when an administrative source is received by Statistics Norway. Methods for editing a single administrative source are more or less the same as for surveys, i.e. linkage to a base register and control of consistency between variables for each unit. This paper concentrates on editing of linked files i.e. linking of files that are edited as single sources. The main challenge in developing integrated systems from a large number of sources is that *cases of inconsistency* are identified when two sources are linked at unit level.

11. The more sources to be linked the more errors are found and corrected for. The correction of an error is part of the editing. The starting point is a through knowledge of the quality of each source that contributes to the linked file. When contradictory information between two sources is identified, it is important to know which of the two sources should be the most reliable. Even the order in which the files are linked affects the final result. In principle, the quality of a linked file should improve when an additional source is integrated. Three challenges in the process of editing and imputation of a linked job file are discussed:

- to identify the same unit of job from different sources;
- to ensure consistency between periods jobs are active when a person changes jobs;
- to develop methods for imputing and calculating variables on the job file.

A. Definition of the unit of job and list of variables in the job file

12. The unit of job is defined within the SNA (System of National Accounts): A job is defined as an explicit or implicit contract between a person and an institution to perform work in return for compensation for a defined period or until further notice. Variables that identify a job:

[1] PIN, BIN, T1 - T2

13. The PIN of the employed person, the BIN of the work place and the period, T1-T2, a job is active identify the unit of job. The concept of job is based on integration of three statistical units, *person*, *work place* and *job*. Examples of variables in the job file listed by the statistical unit are:

person: sex, age, residence, family, educational attainment
establishment: locality, industry, institutional sector, size group
job: occupation, hours paid for, hours actually worked, wage sum of the calendar year

14. The unit of job represents a link between the CPR and the LUR. This link makes it possible to make statistics for the staff of an establishment from variables such as age, sex and educational attainment. Variables related to the establishment such as industry and size group can be specified on statistical files for persons.

B. Data sources for the job file

15. The job file is based on linkage of base registers and a large number of other administrative sources. The process of editing and imputation might affect the operational definition of a variable, for example the classification of a person as employed, unemployed or not in the labour force.

The main administrative data sources for the job file:

Unit of employee jobs:	**A.** Social Security data system on employee jobs **B.** Data system of Tax Agency, where the employer reports the annual wage sum for each employee job
Self-employed jobs:	**C.** Data system of Tax Agency on annual mixed income of self-employed jobs
Spells of unemployment and periods in labour market measures:	**D.** Data systems of Employment Service

16. In source A the BIN of the unit of *establishment* identifies the work place. In source B and C the BIN of the *enterprise* identifies the work place. Most enterprises comprise only one establishment. For the minority of enterprises that are profiled into more than one establishment, Statistics Norway has to control that the enterprise reports to source A are in accordance with the agreement on identifying the unit of job by profiled establishments. The controls on the employer reporting are mostly based on computerized procedures. The follow up of errors in the employer reporting are based on contact with the enterprise. The unit of establishment is a statistical concept and this is an example that Statistics Norway has succeeded to have a unit for mainly statistical purposes implemented in administrative data systems such as the LUR (base Legal Units Register) and source A.

17. During the year 2001 source A identifies *2,903,000* employee jobs. An employer reports employee jobs with a working week of 4 hours or more and jobs with a continuous duration of 6 days or more. The employer reports annual wage sum for *4,331,000* employee jobs to source B. A large number of these are very small jobs that are not reported to source A. The number of annual tax returns for self-employed jobs of source C are *330,000*. The number of spells of unemployment and periods in labour market measure reported to Employment Service during 2001 are *763,000*. Some of the jobs reported to A and B refer to persons not resident in Norway.

18. The task is to integrate these 8,427,000 units into *one* integrated and consistent statistical micro file. The number of persons involved are 2,500,000, i.e. in average there are 3.4 units registered for each person. Most persons are registered with one unit in A and one in B

C. To identify the unit of employee job across sources

First step, Linkage of A and B

19. The result for the year 2001: about 90 % of the jobs in A are linked to a unit in B, i.e. a Social Security job is registered with compensation or labour income in the Tax register. One reason for not finding a job in source B is that the same job is identified with different BIN in the two sources.

Second step, identifying more linked jobs of A and B

20. About 7% of the jobs of A without linkage to B are reported to source B with a BIN for another enterprise than that reported to A. A procedure to identify these jobs and find the enterprise unit used in source A has been developed. The design of the procedure is based on empirical studies of jobs registered in B without a match to a job in A. Usually the difference in the unit of enterprise used in reporting to A and B refers to the whole staff of an employer. There are other sources from labour income than B and the final result of the identification procedure is that 97 % of the jobs in source A have information on the wage sum for the year 2001.

21. The remaining 3% of employee jobs of source A are registered without compensation (labour income) in year 2001 and classified as not in active employment in 2001, i.e. the date of termination of these jobs are corrected to 31 December 2000 or earlier.

Third step, identification of employee jobs registered in B only

22. When the identification procedure is fulfilled there is an annual wage sum registered in source B without match to a job in source A for *1 428 000* employee jobs. Some of these wage sums refer to employment performed in the year 2001. A large number of the employee jobs without match to a job in source A refer to a very small wage sum. To classify a person as employed during the year in Census 2001 there is a limit of 100 hours paid for. This equal an annual wage sum of about 15 000 NOK (2 600 USD). About 250 000 persons are classified as employed on the basis of the 1 428 000 employee jobs without match to a job in source A.

23. For employee in enterprises with two or more establishments the establishment of the job is selected by a statistical model based on variables of residence and information on staff of the same enterprise that live in the same municipality.

D. Methods for ensuring consistency in the dates of start and termination of a job

24. There are some delays in employer's reporting of start and termination of a job to source A, and more delays for termination than for start. The result of this practice is that according to source A, a person might be registered with two active full-time jobs for the same day.

Fourth step, correcting the period an employee job is active

25. For each person there is an edit procedure to ensure that all full-time jobs registered in source A are without overlap to other jobs for the period a job is active. One problem for employee jobs that are reported only to source B is that the majority of the jobs are reported to be active for the period 1 January - 31 December of the reference year.

Fifth step, linking employee jobs and spells of unemployment and measures

26. To improve the dates of start and termination of the employee jobs the units of source D, spells of unemployment and periods in a labour market measure are linked to the file of employee jobs at the level of person. The result of this linkage is that for some persons, inconsistency between the period of an employee job and the period of unemployment or in a labour market measure becomes visible. The result of this confrontation is correcting of dates that improve the quality both for jobs and spells of unemployment. The editing is based on rules that decide the most reliable source. These rules are complicated and the decision depends on information of delays in reporting.

27. The integration of employee jobs and spells of unemployment means that the LFS concept of *labour force* based on administrative sources is implemented.

E. Linking employee jobs and self-employed jobs

28. Some employed perform a secondary job in parallel to the main job. Some employed perform both employee jobs and self-employed jobs in parallel or move from one type of job to the other during the calendar year.

Sixth step, classification of main jobs

29. For each employed person all employee jobs are classified as a main job or as a secondary job. The same classification is implemented for self-employed jobs.

Seventh step, Linking employee jobs and self-employed jobs

30. For employed that perform both employee and self-employed jobs during a year there is a second round of adjustment of dating of jobs and then there is a classification of which of the main jobs should be the most important - the employed person is classified as self-employed or as an employee.

F. Editing and imputation of hours paid for

31. The process of editing and imputation covers other variables than the variables that identify the unit of job, examples are calculation and imputation of hours paid for each job and imputation of item non-response of occupation of jobs classified as a main job.

32. Hours paid for during the calendar year are calculated for each job. The calculation of hours paid for is based on relation [2]

[2[Wage sum = Hours paid for * wage rate

33. Wage sum is registered in source B. Mixed income of self-employed jobs is registered in source C. Wage rates are based on imputation for groups of jobs.

Imputation of wage rates for homogenous groups of jobs

34. The method for imputing wage rates is based on a set of groups of jobs that are expected to be homogenous with regard to wage rate. So far average wage rate for a group of jobs are used in the calculation of hours paid for. The source for calculation of wage rates is the linked file of A and B. Information on occupation is not utilized in the groupings. In future imputation of wage rates should be based on statistical models and calculation of wage rates should include sources such as micro files for wage statistics and the variable of occupation.

35. Persons with an employee job that is registered as a full-time job in source A are sorted in 160 groups by sex, 4 age groups, 1 digit economic activity and 3 groups of educational attainment. Number of working days of a job is based on date of start and eventually date of termination of the job. Average full-time working week is based on the LFS - 38.5 hours for male and 36.6 for female. The total of employee jobs are allocated to one of the 160 groups and hours paid for are calculated by using the average wage rate for the full-time jobs.

36. The source of hours worked for self-employed jobs is the LFS. Groups based on the LFS, sex, primary industry, other industry, 3 groups of hours paid for, 1-9, 20-29 and 30-. According to the LFS, most self-employed work is full-time.

37. It would be useful to split mixed income of a self-employed job in a labour component and a capital component. A staring point could be to study how the Tax Agency splits mixed income into labour and capital components. The next step would be the development of some kind of statistical model to impute labour income for each self-employed job.

G. Editing and imputation of occupation

38. Occupation of a job is specified in source A (Social Security data system on employee jobs), but not in source B (annual wage sum) or C (self-employed). Information on occupation is missing for 3 groups of the job file: (i) employee jobs registered both in A and B, 1 905,000, for 6.4% of these jobs information of occupation is missing, (ii) for employee jobs based on source B only, 210,000, there is no information on occupation from administrative sources, (iii) for self-employed jobs source C, 152,000, there is no information on occupation. For these 3 groups statistics on occupation is based on imputation at the level of job. The figures refer to occupation for main job of employed per 1st November 2001.

39. The imputation of item non-response of source A is based on information on occupation registered in source A. Imputation of occupation for jobs based on source B and C is based on information in the LFS.

40. The grouping for imputation of occupation of main jobs, response homogeneity group (RGH) model, is common for the 3 sources A, B and C. The grouping is designed to create groups that are homogenous with respect to occupation.

Educational attainment: 6, 4 and 2 digits groups of ISCED

41. When the code for educational attainment of a person indicates study field, 4-6 digits the education is often directed towards a specific occupation or a limited group of occupations. This information is utilized in the RHG grouping. In Norway (and other Nordic countries) there is a database of complete registration of fulfilled educational programmes. The system is based on administrative sources and statistical surveys. The current reporting started in 1970 and is linked to information on educational attainment in the Population Census 1970.

Economic activity: 5 and 2 digits groups of NACE (European version of ISIC)

The source for economic activity is the statistical Business Register

Age: 4 groups for source A and 2 for the LFS

Sex. Grouping by sex is not used for small groups.

42. For each group the distribution function of occupation is specified. 352 occupational groups are specified. The distribution function for the actual group is allocated to main jobs of this group.

43. The method for imputation of occupation should be improved. More sources, such as micro files for wage statistics, and variables such as institutional sector, labour income and or wage rate and country background, should contribute to the grouping. Statistical stochastic models should be developed for the different groups of non-response. The statistics on occupation is based on the distribution function for groups. This method should be replaced by imputation of occupational code for each main job with.

IV. EDITING AND IMPUTATION IN DEVELOPMENT OF THE EXTENDED JOB FILE

44. In 2004 Statistics Norway launched two projects that will be of importance for the development of demographic and social statistics. The first project is related to development of the infrastructure of base registers. A second project is a program to develop a coherent and integrated statistical system on person, family and household. Both projects are a further development of the Population and Housing Census 2001. The extended job file developed for the Census 2001 is to be improved and operated as an annual file. The extended job file includes other time use units such as unemployment and education and income sources such as pensions. The job file and extended job file are sources in statistics on labour,

living conditions and population census and sources for research projects on labour market and living conditions.

45. The extended job file includes units of *time use activities* in addition to paid work such as: spell of unemployment, period in a labour market measure, period in current education and in household work and units of *sources of livelihood* in addition to labour income such as: sickness and unemployment benefit, disabled and old age pension, education grant, student loan and contributions from parents.

46. Integration of these units into a linked file needs editing and imputation. In Norway and the other Nordic countries, information on persons in current education from administrative sources are of good quality and have good coverage. Information of hours used in education is limited to some indicators and has to be based on imputation in a system outlined here. For household work there is some administrative sources that indicate that the person is in full-time household work. Most adults living in a private household perform household work and an integrated system would be based on imputation of hours used in household work. In a satellite to the National Accounts the value of household work are calculated. According to the Time Use Surveys the total hours used in household work are about the same level as hours actually worked in the labour market and should be covered in statistics on living conditions. One important aspect of the integrated statistic is that an activity has to be classified as main or secondary activity and an income source as main or secondary source.

V. METHODS FOR MEASURING THE QUALITY OF A LINKED FILE

47. Methods for measuring the quality should cover base registers, use of ID numbers, the sources received by Statistics Norway and systems of linked files. Work on methods is based on studies of linked files of administrative sources and on linkage of administrative sources and household surveys. For both methods it is not enough to describe gross deviations between two sources. To decide which of the two sources should be the most reliable one has to know the reason why the value of a variable is different in the two sources. A much more systematic development of methods is needed. TQM (Total Quality Management) would be the frame for this work. Some projects are in the pipeline.

48. When a statistical variable is derived from a linked file of base registers and a number of other administrative data systems it would be useful to have information on the quality of the variable. Measurement of the quality of a linked file should be based on information of each source and by linking the integrated file of administrative sources and statistical surveys at the level of person. Information on the quality of the integrated file affects the methods for editing and imputation and it should be of interest to improve the methods to measure quality, and as a result of this information improve the methods for editing and imputation.

VI. EXAMPLES OF RECORD LINKAGE IN ADMINISTRATIVE PROCEDURES

49. Record linkage of administrative data in administrative cases and procedures is under development in government agencies. For administrative use procedures for editing and imputation need to be based on some kind of documentation to confirm changes. Additional information could be collected when inconsistent information within a source or between sources is identified and the client could confirm what should be the correct information.

50. The Norwegian Tax Agency has succeeded in preprinting the annual tax return for the majority of households and person. By use of record linkage of administrative data on income and wealth from employers, banks etc. detailed information on income posts are registered. The proposal from the Tax Agency for the tax return for a person is sent by mail and the person has to confirm the proposed tax return or to make corrections. This system has improved the administrative sources on income and wealth, more detailed information is available.

51. The Government plans to unite the local offices and central agencies of Social Security, Employment Service and municipal Social Service. This reform is aiming to reduce sick leave by

measures to reduce the sickness period, to find measures to reduce the increasing number of persons that receive disabled pension and early retirement pension. Administrative cases of the new unit would have to be based on a linked file close to the extended job file described above. The information would have to be organized as longitudinal data. Improved procedures for editing and imputation compared to similar procedures for official statistics would be necessary at the united office.

References

[1] Svein Nordbotten (1966): A statistical file system, Statistisk Tidskrift No 2.

[2] O. Aukrust and S. Nordbotten (1970): Files of individual data and their potential for social research, Review of Income and Wealth.

[3] Svein Nordbotten (1965): The Effiency of Automatic Detection and Correction of Errors in Individual Observations ….., ISI Proceedings.

[4] Denmark Statistics and Eurostat, (1995): Statistics on Persons in Denmark - A register-based statistical System, Eurostat.

[5] Statistics Sweden (2004): Registerstatistik - administrative data for statistiska syften.

[6] Frank Linder (2003): The Dutch Virtual Census 2001.
(Paper for WS on Data Integration and Record matching, Vienna 2003).

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