



Statistisk sentralbyrå
Statistics Norway

e-mail: sga@ssb.no

Record linking of base registers and other administrative data systems - problems and methods

Svein Gåsemyr

Contents

1. Introduction	2
2. Infrastructure for record linking	2
3. Record linking of sources with different reference periods	4
4. Editing and imputation of a system of linked files	5
5. The process of record linking and definition of variables	7
6. Record linking in operating of a set of decentralized longitudinal databases	9
7. Methods to measure the quality of a system of linked files	10
8. Record linking for administrative procedures and casework	12
References	13

<p>Paper for the Siena Group on Social Statistics, meeting in Finland 2005 Session on Record-linking</p>
--

Record linking of base registers and other administrative data systems - problems and methods

1. Introduction

The advantage of using administrative data as sources for statistics is threefold:

- usually the total population of a unit is covered
- events and periods for which the value of a variable is valid are dated
- the costs of transferring extracts of administrative data systems to the NSI are small

The advantage of record linkage is due to the fact that the *value of a data file* as a source for official statistics increases:

- when linked to another source and
- when linked to a the time series from the same source.

Unique ID numbers are used in statistical surveys and usually a statistical micro file is a linked file of the survey, 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 linking of administrative sources. Linking of statistical surveys and registers is an important aspect in developing integrated statistical systems, but the methodological challenges in combined use of surveys and registers are related to differences in definitions of variables and methods of estimation - and not to methods of record linking.

The medium term objective is to develop a coherent and integrated statistical system of person, family and household or welfare accounts. The system should describe the status of the population's living conditions and the changes that have led to the status. The system would be the basis for analysis of different efforts aimed at influencing the living conditions. Efficient methods for record linking of all available sources are a prerequisite for developing a coherent and integrated system at unit level.

The main challenge in developing integrated systems of a large number of sources is that *cases of inconsistency* are identified when two sources are linked at unit level. There can be several reasons why two sources display contradictory information.

This paper presents the Norwegian infrastructure for record linking, section 2, and problems and methods in linking by presenting examples related to: differences in reference periods, section 3, editing and imputation of linked files section 4, definition of variables based on linked files section 5, linking of longitudinal databases section 6, and measuring the quality of linked files section 7, section 8 describes record linking for administrative purposes.

2. Infrastructure for record linking

The development and use of computerized administrative data systems in the private sector and government agencies since its start in the 1960's is based on the same infrastructure and strategy of electronic data collection in all Nordic countries:

- Operation of computerized central base administrative registers for the total population of the most important units, **i)** person, **ii)** business and **iii)** land property/building/dwelling/address
- Assignment of an official and unique ID number for the main units, PIN (Person Identification Number) BIN (Businesses) and DIN (Dwelling) and use of these unique ID numbers in administrative data systems in government and the private sector and statistical surveys

The reference list at the end of this paper provides documentation on the infrastructure for electronic data collection and reuse of administrative data.

The task of the base registers is to identify total populations of defined units. The task of the ID numbers is to ensure reliable and efficient linkage at unit level. The data collection strategy of the government and private sector is that a variable should be reported only to one government agency. A prerequisite for this strategy is reliable linkage at unit level.

There are statistical versions of the 3 base registers in Statistics Norway (SSB). A project to improve the coordination, integration and IT solutions for the 3 statistical base registers started in 2004.

A prerequisite to develop systems of linked files is that the same definitions of units and population are applied in all sources. There is no problem related to the definition of person. The CPR is based on *de jure* residence and not *de facto* residence. Students and spouses living in institutions are two groups that are affected by the practice of CPR. This results in some cases of inconsistency when the CPR is linked to other sources.

The CPR includes residents and non-residents that are registered in administrative data systems of government agencies. The majority of persons registered in administrative data systems of government agencies are registered as residents in Norway, but a substantial number of the CPR population are resident in other countries, examples are foreigners that are employed, are registered with old age pension or other income, wealth or property in Norway.

Economic statistics are based on enterprises registered in Norwegian administrative systems. Some of the employees of these enterprises are residents of other countries, but they are included in the employment statistics. Household surveys and Population Censuses are limited to resident persons. The sample of the LFS (Labour Force Survey) refers to residents. The jobs of respondents of the LFS that work for an enterprise registered in an other country are covered by the LFS.

In the Census and other social statistics the most suitable unit to describe the work place is the establishment. The establishment is a statistical concept and the most important business units for statistics on production and employment. Most enterprises (or legal

economic units) comprise only one establishment. But a large enterprise often has to be profiled in establishments.

The administrative Legal Unit Register (LUR) specifies both the unit of enterprise and the statistical unit of establishment. Statistics Norway is responsible for profiling an enterprise in establishments for the LUR. Of importance for register-based employment statistics and the Census is the fact that when employers report employee jobs to Social Security, the PIN of the employee and the BIN of the establishment identify the same unit of job.

3. Record linking of sources with different reference periods

One of the advantages of variables from linked files of base registers and other administrative data is that the value of a variable is specified by the period the actual value is valid. The reference period can be specified by:

- a date (examples are demographic events such as date of birth and date of start and stop of the unit of job)
- a week (an example is the LFS)
- a month, (an example would be pension and other transfers from Social Security)
- a calendar year (examples would be annual income and annual income tax).

The Census concepts of *usual* activity and *current* activity should be defined operationally and used in register-based censuses and the system for social statistics. A short reference period such as a day or a week is of importance in measuring turning points in production, turnover and in the labour market. Statistics on living conditions, income distribution, poverty and groups that are marginal to the labour market need a long reference period. For some statistics a reference period of more than one year is needed to present interesting statistics. When the task is to describe the dynamics of the labour market and the flows between the main activities of a person a short reference period is needed.

The following is an example of a concept that utilizes the dating of events on the labour market: A person is a full-time member of the labour force for the whole year. The distribution of the labour force into employment and unemployment is described by volume, e.g. 10% unemployed during the year. The ILO definitions for the LFS are based on a short-term period of one week. When the reference period is one year it is of importance to create concepts that describe the realities during the year, i.e. the dynamics of the main statuses or changes during the year.

The ideal solution is to create detailed information on the start and end of the value of a variable. In principle, there could be three methods of harmonization of reference periods:

One of the linked sources gives detailed information on periods,

The population of employee jobs is based on two sources, Social Security data and annual reports on wage sums. The last source very often refers to 1.01 - 31.12. For these jobs the periods registered by Social Security should be the most reliable.

Link an additional source with detailed information on periods

The annual income for a person registered in the data system of income and tax refers to the calendar year. If it is necessary to split the annual income into months, sources with

information that refers to month and day can be linked to the income file, i.e. data on labour income, unemployment benefits and pension.

Imputing detailed information about periods

Self-employed jobs refer to the calendar year. If more detailed reference periods are needed the periods could be imputed, e.g. based on statistical models and data from other administrative sources on income and time use activity and the LFS.

4. Editing and imputation of a system of linked files

Records from statistical surveys and administrative data that are received at the NSI are controlled and corrected if necessary. Methods for editing and imputing a single source are presented in [7] (Statistics Sweden).

4.1 The job file

The interest in this paper is to discuss methods for editing and imputation that are related to processes of record linking. When two sources are linked at unit level one always finds some cases of contradictory information. In developing methods to ensure consistency for all units of the linked file, the starting point would be a thorough study of the reliability of each source. Methods for editing and imputation that are related to record linkage are discussed by presenting methods that are developed for creating an integrated file of:

- employee jobs
- self-employment jobs
- spells of unemployment
- periods in labour market measures.

Variables that identify a job:

[1] PIN, BIN, T1 - T2

The PIN of the employed person, the BIN of the work place and the period a job is active, T1 - T2, 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 that are related to the unit of:

person: sex, age, residence, family, educational attainment

establishment: locality, industry, institutional sector, size group

job: occupation, hours paid for, hours actually worked, wage sum for the calendar year

Three challenges in the process of editing and imputation in a linked job file are discussed:

- to identify the same unit of job across sources
- to ensure consistency in date of start and termination of jobs
- to develop methods for calculation and imputation of variables

Data sources for the linked job file:

- A. Social Security data system of employee jobs
- B. Tax Agency data system of annual wage sum of employee jobs
- C. Tax Agency data system of mixed income for self-employed jobs
- D. Employment Service data system of spells of unemployment and period in a labour market measure

Two qualities of sources A should be noted. *First*, in A an enterprise is split into establishments according to the statistical Business Register. Only a few and usually large enterprises are profiled into establishments. The unit of establishment is a statistical concept, but Statistics Norway has succeeded in having this unit registered in the administrative Legal Unit Register and used to identify the job in the administrative source A. *Second*, in source A the information on date(s) of start and, if appropriate end of a job is reliable in most cases.

Methods for identifying the unit of employee job across sources

The first step is linking of A and B to identify the same job in A and B. There is no problem related to the use of PIN in the two sources. Source B identifies the job by the unit of enterprise. When an enterprise is split into more than one establishment, source A gives the information on the establishments. The main problem in this step is that some employers report an employee job to B by using an ID for the enterprise that differs from the enterprise ID reported to A. A rather complicated procedure is developed to find the correct enterprise unit that identifies the job in both source A and B.

To classify a job of source A as an active unit of employment during the reference year, an annual wage sum for this job has to be reported to source B. Some employee jobs are reported only to the source B. Usually these are small casual jobs that are under the limit for reporting to source A. Some wage sums reported for year T refers to employment performed in year T-1.

The second step is to link employee jobs and spells of unemployment and labour market measures, i.e. link D to the linked file of A and B.

Methods to ensure consistent dating of jobs

During a year a substantial share of the labour force change jobs. Some employed have a secondary job in parallel with the main job. Some persons experience spell(s) of unemployment and/or periods in labour market measures during a year. The third step is to ensure that periods for all active jobs of a person are consistent. An employed cannot be registered with two full-time jobs in parallel at the same date and there has to be correspondence between periods as full-time employed and full-time unemployed. Some dates of start and end are adjusted in this step. A large number of the employee jobs of source B are registered for the period 1 January - 31 December. Some of these dates have to be corrected.

The fourth step is to link source C to the linked file of A, B and D. Information about the period a self-employed is actively employed during the year is missing in source C, but most self-employed jobs are active the whole year. The fifth step is to classify a job as a main or secondary job. To classify a job that is active during the year as a main job there is a limit of 100 hours paid for in the census 2001. This equals an annual wage sum of about 15 000 NOK (2 000 Euro).

Imputation

A large number of employee jobs registered in the data systems of the Tax Agency are dated with start 01.01. and end 13.12. For many jobs this is not really the active period. It is of interest to have a reasonably good quality for the period a job is active for all main jobs. Most persons have one or a few jobs during a year and it is easy to specify changes during the year. For more complicated cases, jobs, spells of unemployment, current education etc., the process of linking the sources includes methods for ensuring consistency between different time use activities during the year. It might be a solution to use statistical models to impute the period of time use activities.

Methods of editing and imputation are related. Some of the errors identified in the linking of two sources are corrected for by imputation. Imputation of item non-response mass-imputation for a variable for the total population is sometimes necessary in developing a system of linked files. Examples of imputations in the job file are occupation of main employee job, (for about 16% of employee jobs information on occupation is missing), and calculation of hours paid for.

4.2 The extended job file

The extended job file covers main time use activities and sources of livelihood. Time use activities that are to be included, in addition to labour force, are current education and household work. Income sources are labour income, transfers, student loan and grant and the calculated value of household work.

The source for imputing time use in current education is the data system of persons in current education. There are some administrative sources indicating that a person is in full-time household work. The sources for the statistical model to impute time use in current education and household work should be linked to files from Time Use Surveys and administrative sources. The methods of imputation should cover both social statistics and National Accounts.

The extended job file has a key role in the development of coherent and integrated social statistics and in integration of economic and social statistics. Statistics Sweden classifies the extended job file, ("the activity register") as a fourth base register. The reason for this classification is the important role the extended job file has in the process of record linking and integration of sources. The unit of job represents the link between the CPR and BR (Business Register). Like the other 3 base registers some variables are direct related to the unit of job. It is not necessary or useful to introduce a specific ID number for the unit of job. For this reason Statistics Norway does not see the extended job file as a base register.

5. The process of record linking and definition of variables

Variables to be collected by household surveys are limited to what a respondent is able to report. Business surveys are usually limited to information that is easily available from internal administrative systems of enterprises. Administrative data systems comprise variables that are needed in administrative case-works and procedures. Some statistical concepts are needed for administrative purposes and Statistics Norway has succeeded in having statistical concepts included in administrative data systems. Statistics Norway is responsible for some activities in the operation of administrative data systems such as profiling enterprises in establishments, coding economic activity of establishments and

occupation for employee jobs. Statistical standards on health and education are implemented in administrative data systems.

The experiences gained are that registration of statistical units and variables in administrative sources can be very difficult. Statistics Norway has to implement measures to ensure that such units and variables are used in accordance with agreed definitions etc. The unit of establishment is registered in the Social Security register of employee jobs. As there is no administrative need for this unit Statistics Norway has to control that the involved employers use the correct unit of establishment in reporting to Social Security. A computerized system selects enterprises for manual control and contact with the enterprise.

The infrastructure of base registers and the use of official and unique ID numbers are developed to promote efficient data collection for administrative and statistical purposes. One of the principles to ensure efficient data collection is that the same variable should be reported to a government agency only once. This means that a household survey should not include a variable that is available from administrative data. There are some exceptions to this principle:

- Variables to be used to measure the quality of administrative data
- Variables that are needed for short-term statistics

Examples of units and variables collected both in surveys and administrative sources are job, income and household composition.

Variables that are derived from linked files of administrative sources should be an important tool in creating integrated statistics. Sometimes the linked sources are related to different statistical domains. Examples of such variables are time use activities and sources of livelihood.

Labour Force - job, spell of unemployment, labour market measures

The LFS is designed to measure turning points on the labour market. The concept of labour force, whose members are either in a job or actively in seeking a job, represents the supply of labour. The demand of labour is the sum of filled posts (jobs) and unfilled posts. The labour force concept of register-based statistics should include labour market measure since this unit unquestionably represents the supply of labour and plays an important role in labour market policy. The time use activities of labour market measures comprises ordinary jobs, labour rehabilitation, (medical rehabilitation is not included in the labour force), and current education. Labour market measures are not included in the ILO definition for the LFS as this concept is too complicated to be based on a block of interview questions.

Eurostat has developed a set of core variables for social statistics. Core variables are variables that are used across statistical domains and are therefore an important tool for integration of statistics. The Eurostat definitions are related to household surveys. Statistics Norway has started a project to develop a national set of core variables. Most Norwegian core variables will be based on base registers and other administrative data systems.

The best practice to ensure a uniform implementation of harmonized core variables in various domains would be to organize a common database for all core variables. As all statistical micro files of the unit of person are linked to the statistical CPR, an efficient solution could be to include all core variables in the statistical CPR.

6. Record linking in operating of a set of decentralized longitudinal databases

For most statistical micro files the unit are identified with the common ID number (or encrypted systems) and can be linked across sources and through time. The time-series for some variables starts in 1960. For each year since 1960 more and more variables are covered in stored files. Only a small fraction of these data are utilized. More efficient procedures that improve the access to the stored micro files are needed. This year a large project has been launched to develop an improved infrastructure for using statistical micro files in research projects. Step by step systems of longitudinal data- bases are developed. About 10 systems are in operation or under development.

Statistical base registers on person, business and dwelling are organized as longitudinal databases. A project to develop these registers in a coordinated way started last year. Other important systems in operation are: database on education, income accounts, transfer, labour market. There are data sources for health indicators, annual census files, and nursing and care.

The strategy in Norway and other Nordic countries is to develop a decentralized system of databases and an efficient menu-based way for linking the bases at unit level.

6.1 Longitudinal database on detailed information on National Insurance data

Statistics Norway operates a longitudinal database on disabled pension, refusal of application for disabled pension, early old age pension, old age pension, survivors' pension, family allowances, sickness allowance, social assistance and a number of temporary transfers. Operation of the database started in 2000. The time series start in 1992. The database comprises longitudinal background variables on demography, education, labour and income and wealth. There are no controls on consistency *between* the linked sources.

The annual maintenance is based on record linkage of base registers and other administrative sources. The database is the source of many research projects every year. The data model and IT solution need to be further developed.

6.2 A decentralized system of about 10 longitudinal databases

The plan is to develop a system of about 10 independent databases that are coordinated and organized in a way that would make it easy to select the necessary units and variables for a specified period. The record linking is to be organized by a menu.

The 3 statistical base registers are organized as longitudinal databases.

The statistical CPR - some time series start in 1960

The statistical Business Register - some industries start in 1956, the coverage of all industries starts in 1995.

The statistical base register on land property, building, dwelling and address - the last component, the unit of dwelling, is operated from 2001. The Housing Census 2001 contributed to the initial file of the Dwelling register.

A project to improve the coordination, linking at unit level and IT solutions for the 3 base registers started in 2004.

Other bases in operation and under development are:

- Education - the initial file is based on the 1970 Census
- Income accounts - some sources starts in 1967
- Labour market - detailed information from 1992, in use for research project
- Annual micro census file
- Nursing and care - planned to be in operation within 2-3 years.

There are sources available to develop a database on health indicators. Some of the health data are very sensitive and for the reason of privacy separate files for these data has to be created and there are rules for a project that needs access to some of the variables.

A project to improve the infrastructure for creating micro files for research projects starts this year. Methods of linking existing statistical micro files, longitudinal data and other files, and IT solutions for linking would be important components in all these projects. The need for better and more standardized documentation of units, variables, editing and imputation becomes a key question when a large number of sources are to be linked.

In the process for operating statistical systems some control of consistency of units and variables *between* the databases would be made, i.e. in creating core variables that are derived from sources across databases. In some research projects the researchers want to be responsible for the methods on editing and imputation that ensure consistency, but usually research projects are based on the final version of a statistical file.

7. Methods to measure the quality of a source and a system of linked files

The result of an advanced infrastructure for collecting and linking a large number of sources at micro level is that errors and contradictory information are identified. The more sources to be linked the more errors are found and corrected for. The correction of an error is a part of the editing. The starting point is a through knowledge of the quality of each source. 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.

Methods to measure 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 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.

Use of official and unique ID numbers

The use of official ID numbers by all private and government institutions and private households is the main tool to ensure reliable systems of linked files. The ID numbers comprise one or two control digits to ensure correct typing of an ID and there is a control of the ID when an administrative file or a survey file is received at Statistics Norway. The second step is usually a linking of the received file to the actual base register. The aim of

this linkage is to ensure that the registered unit is an active one and for transfer of variables from the base register.

Examples of linkage to other administrative sources

Some persons that are classified as resident in Norway according to the CPR are not found in any administrative data system e.g. during the last calendar year. This is an indication that the person has emigrated from Norway without reporting this to the population registration.

The Tax Agency is responsible for the population registration and CPR and has asked Statistics Norway for a project to measure the quality of the CPR. Statistics Finland carries out this kind of measurement regularly. Methods for a Norwegian project are outlined and based on linking the CPR with the LFS.

Linkage of registers and household surveys

A new register-based statistics on couple, family and household is under development. Because of delays in reporting to the CPR and use of imputation to group a cohabiting couple more systematic information is needed on the quality of CPR than what is available today. A project to use the LFS to measure the quality of CPR is in progress. Variables to be measured in the LFS and compared to the CPR are residence, household composition, dwelling and couple.

Some experiences have been gained in studying linked file of the LFS and administrative sources on jobs and spells of unemployment. There are proposals for more systematic and through studies of a system of linked file and the LFS sample.

Checking overall consistency

A set of register-based core variables to be used in all statistics on persons and households should be the main instrument to ensure a statistical system with overall consistency. Statistics Norway does not apply a policy of one number. Minor biases and differences e.g. between short-term statistics and final structural statistics or between social and economic statistics, would be accepted. If an error in a register-based variables is significant, e.g. because of delays in reporting, it might be useful to adjust at macro level.

The idea is to develop a decentralized system of about 10 integrated and large databases organized as longitudinal data. For each of these domains Statistics Accounts should be developed. Some of these accounts are already in operation examples are accounts on demography, education, income and labour. A Welfare Account (WA), covering most of the domain of social statistics should be the framework for the accounts on statistical domains. The plan is to start the developing of a WA by studying statistics on life path and time series of cohorts. This means record linking across domains and through time.

Adjustment for measurements errors in a system of linked files

The ambitions in developing register-based statistics are to publish pure register statistics on demographic, education, labour market, labour income, transfer, census etc. A prerequisite to achieve this is that the the data sources are of sufficient quality. When the quality of a register-based statistics is found to be insufficient for publishing it might be a solution to adjust the register-based statistics at macro level. The source for macro adjustments would be household surveys.

Administrative data for short-term and temporary statistics

The developing of an annual Census micro file is an important framework for the integrated statistical system on person family and household. One of the problems related to using the census file is the fact that the final version of the Census file for the calendar year T is ready at the end of year T+1. A temporary Census file for year T is needed in April year T+1.

Important short-term statistics on persons and jobs are the LFS and administrative sources on demography and labour market. The LFS statistics is based on combined use of the sample survey, base registers and other administrative data.

8. Record linking for administrative procedures and casework

Record linkage for statistical purposes at unit level and based on computerized procedures were an important aspect when planning the establishment of the infrastructure of base registers from the very beginning, i.e. in the 1960s.

Record linkage of administrative data in administrative case works and procedures are under strong development by government agencies. For administrative use, editing and imputation procedures need to be based on some kind of documentation when information is corrected. 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.

The Norwegian Tax Agency has succeeded in preprinting the annual tax return for the majority of households and persons. 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 make corrections. This system has improved the quality of administrative sources on income and wealth and more detailed information is available.

The Government plans to unite the local offices and central agencies of Social Security, Employment Service and municipal Social Service. This reform aims to reduce sick leave by measures to reduce the period of absence from work and to create measures to reduce the increasing number of persons that receive disabled pension and early retirement pension. Administrative case works of the new unit would have to be based on a linked file with a content of variables that are close to the extended job file described earlier. The information has to be organized as longitudinal data. The methods to be used to edit linked files for administrative use need to be improved compared to methods that are in operation for statistical files today.

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] Statistics Finland (2004): Use of Registers and Administrative Data Sources for Statistical Purposes - Best Practices of Statistics Finland, Handbooks 45
- [4] Denmark Statistics (1995): Statistics on Persons in Denmark - A register-based statistical System, Eurostat
- [5] Denmark Statistics (2000): Register-strategien I det personstatistiske system
- [6] Statistics Sweden (2001): The future development of the Swedish register system
- [7] Statistics Sweden (2004): Registerstatistik - administrative data for statistiska syften
- [8] Van Tuinen et al (1994): Surveys, registers and integration in social statistics, Statistical Journal vol 14 no. 4
- [9] Netherlands Official Statistics (2000): Integrating administrative registers and household surveys. Vol 15 Summer 2000
- [10] Frank Linder (2003): The Dutch Virtual Census 2001
(Paper for WS on Data Integration and Record matching, Vienna 2003)