Session 4

Seminar on Registers in Statistics - methodology and quality
21 - 23 May, 2007 Helsinki

A Norwegian set of harmonized core variable based on registers - definitions, problems, quality and methods

By

Svein Gåsemyr, sga@ssb.no, and Johan-Kristian Tønder, jkt@ssb.no, Statistics Norway

1. Introduction

The task of a NSI is to develop a coherent and integrated statistical system. An integrated statistical system ensures that the national users get statistics that are comparative across domains. Some users of statistics need statistics that is comparable across domains and countries. International organizations such as UN, OECD and EU work out recommendations on statistics aiming to harmonize official statistics across domains and countries. The EU even adopts statistical acts for the most important statistics that are needed for European policy making.

The European work on harmonization of social statistics started with the London Workshop November 1996 and was followed by Task Force and Working Groups meetings. The proposal is to introduce systematically a set of core harmonized variables in each social survey (or data gathering through registers) involving transmission of micro data to Eurostat.

The inclusion of a limited set of core variable in all European household surveys is done for analytical purposes, as background variables and not to measure the phenomenon. The functioning of a core variable is to be used to group the population covered in a statistical table in subpopulations of interest. A set of harmonized core variables is introduced in surveys to ensure comparable statistics across domains and across countries.

This year Eurostat has approved a set of harmonized core variables for social statistics. The Eurostat recommendation focuses on definitions of core social variables to be used in European household surveys such as the LFS, EU-SILC, HBS and Census. For some member states the European household surveys are based on a combined use of **i)** household interviews, **ii**) base registers and **iii**) other administrative data systems. The Eurostat documents have some references to register based variables, bur do not discuss problems related to use of register based core variables.

Most of the harmonized core variables recommended by Eurostat are register variables in the Norwegian (and Nordic) statistics. The paper present a proposal for a Norwegian set of core social variables, data sources (base registers and administrative data systems), definitions and problems related to comparison of variables based on registers and the same variable based on interview of persons. For Statistics Norway it would be useful to work on core social variables and core economic variables in parallel. Some core variables are common for economic and social statistics, and will be discussed as this in the paper.

Data sources for register based social core variables
Most register based social core variables are collected from 3 administrative and statistical *base registers* on person, business and address/dwelling or on a linked file of 2 base registers. The register based *job file* is the source for socio-economic information. The job file is an important tool to integrate economic and social statistics. For this reason Statistics Sweden use the term the fourth base register on the extended job file or activity register, i.e. a linked file of the units of job, spell of unemployment and in current education.

Other sources are the database on fulfilled educational programs (educational attainment is a derived variable), and the database on income accounts. The annual micro file for population and housing census should be the source for the core variable: self-declared labour status. Self-declared labour status would be a derived variable from a linked file of many administrative sources.

Some core variables would be based on linked files of base registers and other administrative data and are a kind of derived variables of several sources. Some register variables have to be based on mass imputation for subgroups. For European household surveys register based imputed variables can be replaced by survey variables also for the Norwegian version. But when core variables are used in register based statistics imputed core variables for subgroups are needed. This paper presents some initial ideas in developing a set of register based core variables. One question to be evaluated is how far we can go in using the method of mass imputation to ensure a complete set of register based core social variables.

2. Definition of variables and role of statistical standards

The traditional tool to integrate statistics across domains is statistical standards. There are different types of standards: **i)** statistical systems (example: System of National Accounts), **ii)** statistical groupings, (examples: large standard classifications such as economic activity and occupation and minor groupings such as martial status, and **iii)** statistical variables, (examples: employment, income and turnover).

Definition of variables
The definition of a concept is independent of the data source, a household survey or an administrative source. Example: definition of employment according to the SNA. The SNA define an employee job and a self-employed job by legal relations. The operational definition of a variable is related to the source. In the LFS a person is classified as employed or not by use of a block of interview questions according to the ILO definitions for the LFS. In register based statistics a person is classified as employed or not, on the base of a job file that is a linked file of 2 base registers and several administrative data systems in Tax and Social Security agencies. The job file is based on the unit of job.

Implementation of variables
A survey core variable is registered in the micro file of the statistical survey on the basis of a block of interview questions. A register based core variable is collected from a statistical database, in most cases from one of three statistical base registers. Statistical micro files for surveys and register based statistics have different reference periods. A database for register based core variables to serve different reference periods should be organized as a longitudinal database. One advantage of the register solution is that a core variable for a given reference period and person will not only have the same definition across domains. The core variable

will have the same value across statistical domains and for the time period of which the value of a variable is valid.

The National Accounts and core social variables as tools for integrated statistics
The NA has been in operation for more that 50 years and has served as an important tool for developing coherent and integrated economic statistics. A set of core social variables could be seen as a first step towards an integrated system of social statistics. The NA is limited to some few aggregated indicators, e.g. total income for the household sector. In social analysis statistics on income distribution and on the existence of household that are facing hardship are needed.

3. The Eurostat recommendation on Core Social Variable

Most core variables are used in a large number of tables and across statistical domains. All statistics are aiming to create groups that are homogeneous in one or more dimensions. The core variables are selected to create important homogenous groups of persons, households and establishments. The systematic inclusion of the set of core variables in every survey or register based data collection will deepen the household or individual analysis by taking into consideration the interaction between a limited number of socio-demographic variables.

3.1 The Eurostat set of core social variables

To ensure a complete list of register based core social variables subgroups for some core variable have to be based on imputation. These core variables are marked by *.

Demographic information
- [1] Sex
- [2] Age in completed years
- [3] Country of birth
- [4] Country of citizenship at time of data collection
- [5] Legal marital status
- [6] De facto martial status (consensual union)*
- [7] Household composition*

Geographic information
- [8] Country of residence
- [9] Region of residence
- [10] Degree of urbanization

Socio-economic information
- [11] Self-declared labour status*
- [12] Status in employment
- [13] Occupation in employment
- [14] Economic sector in employment (i.e. industry)
- [15] Highest level of education completed
- [16] Net monthly income of the household

Some variables are to be added later to be able to construct [17] A European Socio Economic Classification, i.e. size group of work place and responsibility/tasks for some supervisory occupations.

4. A Norwegian list of core social variables

The Norwegian sources for European core variables register also other background variables that should be included in a Norwegian list of core social variables. A Norwegian set of core social variables should include:

Dependent child*

Institutional sector

[18] Size of work place
[19] Responsibility/tasks of some supervisory occupations*
[17] Socio Economic Classification

Main source of livelihood
Main time use activity
Main social status (Self-declared labour status)

Indicators on housing

Indicators on health

4.1 Possible core variables for economic statistics

Some of the core social variables are background variables of economic statistics as well. A Norwegian system of core variables should cover both social and economic statistics. The definition of a social and economic variable might differs for some statistics, but it should be useful to limit the number of different definitions for the same concept.

Core social variables that also are core economic variables

Region, residence of households and establishments
Labour status
Status in employment
Occupation
Highest level of education completed*
Income
Economic activity

Institutional sector
Size of work place
Responsibility/tasks of some supervisory occupations*

Other core variables for business statistics

Economic units by legal type
Demographic variables of establishment and enterprises

Type of unit by type of competition from foreign countries
Input and output products and services

Core variables in National Accounts

Output
Intermediate consumption
Value added
Compensation of employees
Household consumption expenditure
Government consumption expenditure
Gross fixed capital formation
Exports
Import
Gross domestic production
Gross national income
Consumption of fixed capital
Operation surplus

5. Implementation of core social variables by sources

The infrastructure to promote electronic reporting to government agencies and reuse of existing administrative data systems has been developed for about 50 years in Norway, (and in the other Nordic countries). The infrastructure covers administrative and statistical base registers for the most important units, official ID numbers, PIN, (person), BIN, (business) and DIN, (dwelling). The 3 statistical base registers are organized in a common data model and are operated by one common database. The job file should play a key role in a system of harmonized core variables

Variables of the job file:

(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, type of family, educational attainment

establishment: locality, industry, institutional sector, size group

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

The job file is an example where all core variables are integrated in an efficient way. Variables related to the unit of person can be linked to the unit of establishment by the staff of the establishment. Variables related to the unit of person can be transformed to a micro file for unit of establishment by the staff of an establishment. Variables related to the unit of

establishment can be translated to a micro file on persons by the members of the staff. These qualities demonstrate why the job file is an important tool for integration of statistics.

The extended job file should include other activities such as unemployment, in current education and household work and sources of livelihood. For household work we find some indicators from administrative data that the person in question should work full time in household work. There is no general information on household work in administrative data systems. According to Time Use Surveys (TUS) the total number of hours actual worked in domestic task is about the same level as total hours worked in the labour market. Estimates of hours a person use for domestic tasks in a register based micro file have to be based on imputation.

Statistics Norway has succeeded in the efforts to have statistical units and variables registered in administrative base registers and other administrative data system. The statistical unit of establishment is registered in the administrative Legal Unit Register and in the Social Security data system on employee jobs. The large statistical classifications such as NACE, ISCO 88 (COM), ISCED and ICD are used in administrative data systems.

A. Statistical Population Register

The Central Population Register (CPR) includes all persons that have been assigned a PIN. The resident population is the main group of the population. Not-resident persons that have some relation to one or more administrative data systems of government sector, e.g. employed in a Norwegian establishment, and with income from a Norwegian source, are assigned a PIN and registered in the CPR. A government project aiming to measure the group: "illegal residents in Norway", is in progress.

The Nordic countries constitute one population registration region. This means that there is an integrated system of migration statistics for the Nordic region. The statistical CPR is the source for, [**1**] sex, [**2**] age, [**3**] country of birth and, [**4**] citizenship, and [**5**] legal marital status. These variables are related to legal events and should be reported fast and correct.

The problems of using the CPR as a source for statistics are related to *de jure* and *de facto* residence, e.g. students, and delays in reporting family migration. The plan is to use administrative data on *de facto* residence for students and residents living in institutional households as supplementary information within the statistical CPR.

B. Statistical base register on property, address and building, GAB

The geographic variables are registered as a variable of the unit of address, the address identify the unit of dwelling. Geographic variables are: address, [**8**] Country of residence, [**9**] Region, [**10**] Degree of urbanization, and Dwelling ID number, the Dwelling ID number is a part of the numeric address. Indicators on housing are related to the unit of dwelling

Problems related to the GAB registers are delays in reporting, (it is the municipalities that are responsible for reporting to GAB) and for some dwellings there are no residents according to the administrative CPR. Information about the use of these dwellings could be based on a sample survey on "empty" dwellings. Until now, the cadastre register has been organized by

two components the GAB and the digital land property map, DEK in the Norwegian language. At the moment the two components are to be integrated into one common database.

C. Combined use of CPR and dwelling unit of GAB

The core social variables [**6**] De facto marital status (consensual union), and [7] Household composition are based on a combined use of the CPR and GAB. From the linkage of CPR and GAB some dwellings are identified and registered with no resident persons according to the CPR. Information about the reason why a dwelling is empty is needed. Some "empty" dwellings are used as holiday home or occupied by a student that is registered as resident in the parent home.

Household composition
The household concept is the dwelling household, i.e. persons registered in the same dwelling according to the CPR. About 98 percent of the households consist of one-person households or one family nucleus household. For about 5 percent of the families and one-person households the address does not identify the unit of dwelling. This means that an address are related to more than one dwelling according to the GAB and two or more families of the address are not allocated to the dwelling of this address according to the CPR.

Statistics Norway started to publish official register based household statistics in 2006. Statistical methods are used to allocate a family or person to an empty dwelling of the registered address in the statistical CPR. The administrative CPR can not use the method of imputation.

Consensual union
About 15 percent of Norwegian couples live in a consensual union. There is no legal reporting on consensual unions. Consensual unions are created in the statistical CPR by use of rules and a simple statistical model. The starting point is the unit of dwelling household according to the administrative CPR. If two residents of opposite sex have a common child the two constitute a consensual union. If two not related residents of opposite sex and in a suitable age they are imputed as a couple or not according to a statistical model. The model is based on the Census 2001 micro file and Census statistics on consensual unions. There is some updating of the Census micro file from reports on migration, death, changes in legal marital status and other demographic changes to the CPR. When the Census statistics become outdated there should be updated statistics on consensual unions based on household surveys.

The register based statistics on couples, family nucleus and households
The methods of imputing the DIN and consensual unions result in unbiased estimates for the national level. As the missing of DIN is concentrated to Oslo some adjustments at statistical level on regional distributions, Oslo and the rest of the country; have to be made.

D. Statistical Business Register

The BR registers the following core variables:

[**14**] Economic activity (economic sector in employment)
Institutional sector
Size of work place
Economic units by type

Demographic variables of establishment and enterprises

E. Combined use of GAB and BR

The regional variables for the units of enterprise and establishment are collected from GAB when needed by the use of address.

F. Database on start and fulfillment of educational programs

The schools and other educational institutions report, at the level of student, start and fulfillment of educational programs either to an administrative data system of government agencies or direct to Statistics Norway. The statistical database of this reporting is the source of the core variable *in current education.* The core social variable [**15**] *the highest level of education attained* is derived from the database and registered in the database. The starting point of the statistical database on education is the information on educational attainment collected in the Population Census 1970. The reporting on current education started in 1971.

Education fulfilled in foreign countries is missing in the database and statistical surveys on education fulfilled by immigrants are carried out as a part of decennial register based censuses.

G. Database for income accounts

The database for income accounts is based on the concepts and definitions of the Canberra group. The income accounts are based on a linked file of all available administrative data on income and income tax. Some of the sources report to the Tax agency. Some income and source of livelihood is not liable to pay tax such as social transfer and student loan. These data has to be collected from administrative data that do not report to Tax Agency. These data has to be collected from other administrative agencies like Social Security (e.g. family allowances and cash-for-care benefits), The State Educational Loan Fund (scholarship and loans or the State Housing Bank (dwelling support).

There are some studies on the comparison of income distribution when the income variable is based on household surveys and administrative sources. More thorough studies on this topic are needed.

H. Linked files of base registers and administrative data systems

Two large systems of linked files are sources for core variables, the job file and the annual census micro file.

H.1 The database for register based labour market statistics

The units of the database for labour market statistics are; job, spell of unemployment and a period when a person is active in a labour market measure. These 3 units are sub groups of the labour force. The ILO definitions of the Labour Force Survey do not include the concept of persons active in a labour market measure. The main reason for this solution is that it is too difficult to collect information on labour market measure in a household interview. The definition of labour force in register-based statistics must include measures since this is an

important labour market instrument for groups that have problems to get a job in the ordinary labor market. The activity of some labour market measures is education.

The database for labour market statistics is the source of 3 core variables: [**12**] Status in employment, [**13**] Occupation in employment, and the Level of supervisory responsibilities. The last core variable for some occupations is needed for socio-economic groupings and should be based on combined use of the LFS and the register based database for labour statistics. There is no information in administrative sources on supervisory responsibility and this variable has to be based on imputation.

Subgroups of *status in employment:*

- Self-employed
- Employee
  - with permanent job
  - with temporary job

The classification of an employee job as permanent or temporary has to be based on a set of rules that utilize variable on the job file such as the period a job has been active, i.e. a kind of imputation.

Occupation is related to the unit of job. Occupation is registered in the Social Security data system of employee jobs. Statistics Norway is responsible for the coding of occupation in the administrative register according to ISCO 88 (COM). Information about occupation is available from administrative sources for about 85 percent of main employee jobs. The occupation of remaining jobs is to be based on imputation. The imputation model is based on status in employment, economic activity, educational attainment and demographic variables.

H.2 The extended job file as a contribution to the annual census micro file

The annual micro file for the register based population census classify *main source of livelihood* of a person and *main time use activity.* Main social status is a combination of these two variables. The core social variable [**11**] Self-declared labour status is close to the concept of main social status.

The subgroups of Self-declared labour status are:

- Employed
- Unemployed
- In current education, training and unpaid work experience
- In retirement or early retirement
- Permanently disabled
- In compulsory military or community service
- Fulfilling domestic tasks*
- Other inactive persons

I. Database for existing health indicators

Background variables on health status should be the most important background in many statistics and analysis. However, there are no plans to implement a core health variable. Administrative data on health are under strong development in Norway for the time being.

Unfortunately for the time being there are some restrictions to use administrative health data as source for official statistics. We hope that this practice would be changed in the future. A possible health indicator should be based on some of the existing administrative sources such as:

Medical report on birth
Long-term absence from work, unemployment and education for the reason of illness
Medical rehabilitation
Labour market rehabilitation
Application for disabled pension
Refusal or granting the application for disabled pension
Degree of disability
Disabled pension, temporary or permanent
Early retirement
Old age retirement
Use of caring services, at home and in an institution
The causes of death

J. Database for enterprise accounts

In the near future the database on enterprise account will cover all enterprises, also small family companies and farms and enterprises in government sector reports accounts and trading statements. It is expected that this database will result in a strong development of register based economic statistics.

6. The quality of administrative and statistical base registers and core variables

As base registers and important administrative data systems such as the sources for the annual census file are the backbone of the Norwegian official statistics, information on the quality of these sources is needed. At the moment information on quality od administrative data is very limited. We are even missing well developed methods to measure the quality of administrative data.

Statistics Netherlands has proposed a project to develop methods to measure the quality of administrative data as a source for official statistics within the European FP7. Statistics Norway hopes very much that this project would be realized. In any case there is need for a strong international cooperation in developing methods to measure the quality of base registers and other administrative data.

Referanser

(1) To be published by UNECE (2007): Register-based statistics in the Nordic countries – Review of the best practices with focus on population and social statistics.

(2) Eurostat, Danmarks Statistik (1995) Statistics on persons in Denmark: a Register-based Statistical System.

(3) Doc. Eurostat/F/06/DSS/10/13/EN: Final report from the Task Force on Core Social Variables presented at the SDG meeting on 2 February 2007