Working Paper No. 14
ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE/Eurostat Work Session on Registers and Administrative Records for Social and Demographic Statistics
(Geneva, 1 - 3 March 1999)

Session 3, invited paper

Registers in Official Statistics: A Swedish Perspective

Submitted by Statistics Sweden[1]

1. Background

Swedish official statistics is based upon a comprehensive use of registers based on administrative data. The registers are used as a sampling base for traditional surveys, to complement survey data with register information and for pure register statistics, and commonly information is taken from two or more registers.

Despite of having been used in this way for a long time, however, the registers have functioned independently and have not previously been seen as a *comprising statistical register system*. This view has now changed, and today the registers are indeed looked upon as parts of a system - a system constituting a strategically important resource, which need to be further developed and exploited. The present state of affairs at Statistics Sweden can be summarised as follows:

- There are many good registers and several good subsidiary systems, but co-ordination of the system as a whole is needed. Links, reference periods and definitions must be standardised.

- Work on register statistics has led an obscure existence and is poorly documented. There is in many cases a lack of metadata.

- Duplication of work occurs - many integration registers have in part the same content.

- Competition about the same clients has a harmful effect on the climate of co-operation ("My register, my metadata, my clients").

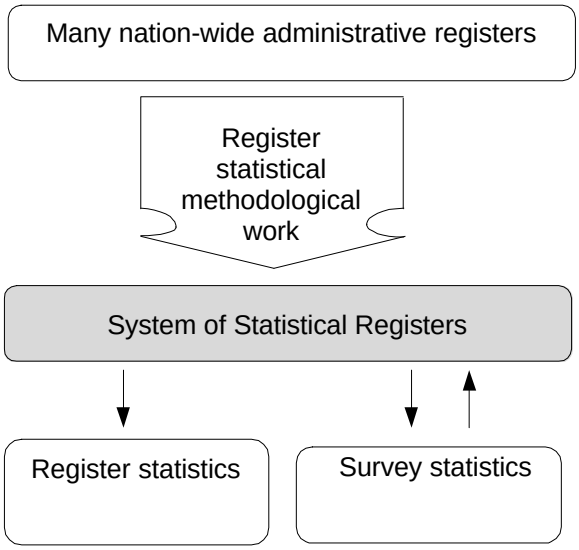
In 1996 began "*The Register Project", its main task being the following: "Statistics Sweden's different registers, which are based on administrative sources, shall be developed into a well-functioning register system. The various parts of the system shall be clearly defined, and by means of increased co-operation the system shall be made more efficient so that new integration registers can be created in order to meet new requirements on the part of the users. Register-statistical methodological work shall be rendered visible and be developed so that the quality of the register statistics can be described and improved*."

[1] Prepared by Evert Blom and Folke Carlsson.

An efficient statistical system should contain both surveys and registers based on administrative data. We wish to stress that register statistics based on administrative data should not be regarded as a cheap, low-quality alternative to survey statistics - they have their own merits and are in many cases the best alternative.

This paper describes the Swedish Register System and the linkage between the different registers. It contains a few examples of register statistics where the links have been utilised and gives an overview of methodological problems. To meet our users' needs we must continue to improve the system, must achieve better linkage, greater consistency and better documentation. We must also develop a common methodology for the statistical work behind a system of statistical registers.

2. The Swedish Register Model



Nation-wide administrative systems differ a lot between countries and the same holds for the possibilities for the NSI's to get access to administrative registers. In Sweden like in the other Nordic countries the use of administrative registers has long been an important basis of the statistical systems. Surveys and registers based on administrative data interact - surveys can be improved by access to registers and register statistics can be evaluated and improved by surveys. Auxiliary information from registers can reduce sampling errors in surveys and survey data can be used to improve the quality of register statistics (see figure beside).

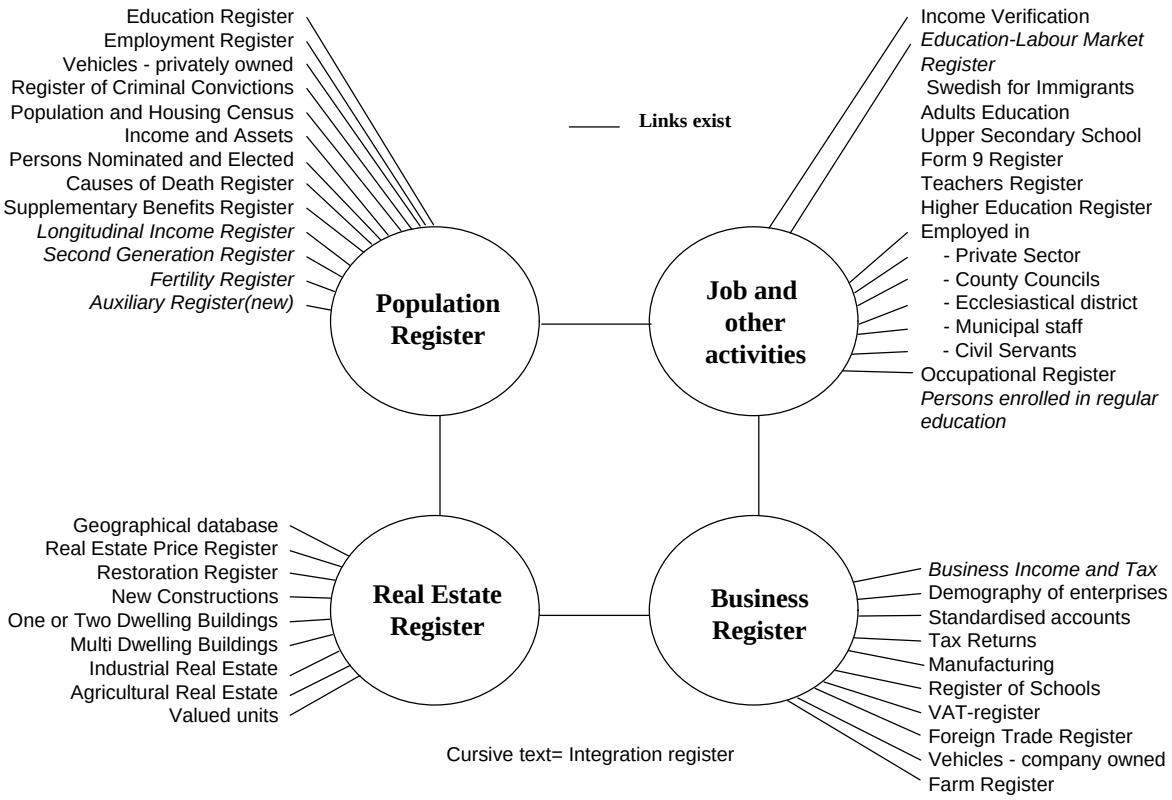
We differentiate between the *administrative registers*, which other authorities submit to Statistics Sweden, and the *statistical registers* which Statistics Sweden has produced from these after, in many cases, extensive *register-statistical methodological work*. The system of statistical registers is the basis of all statistics produced by Statistics Sweden. In all sample surveys registers are used as sampling frames, and register variables are used instead of asking the respondents and also as auxiliary variables when it comes to the estimation. In addition to these survey statistics Statistics Sweden produces register statistics based directly on the processing of information in our registers. The extent of our register system can be seen from the diagram on the next page, where the most important registers are shown.

A statistical register system is made up of the following parts:

- *Basic registers* - registers of objects of fundamental importance
- *Other registers* - registers which can be set in relation to the object volumes of the basic registers
- *Links* between the objects in different basic registers and between basic registers and other registers
- *Standardised variables* - variables of fundamental importance
- *Metadata*
- *Register-statistical methods* and routines for *quality assurance*

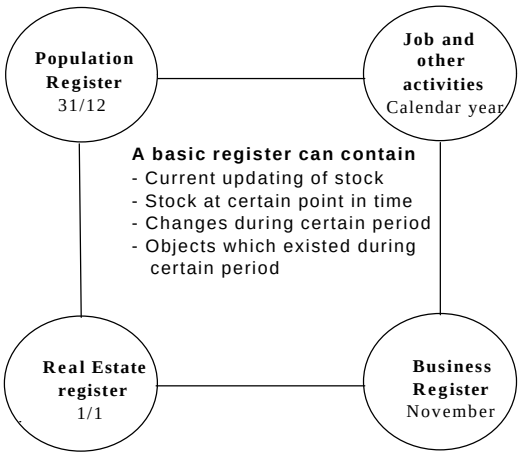
We have defined four basic registers for four types of object and made an inventory of other important registers connected with these types of object. By means of different diagrams of the system we have illustrated a number of problems which have been discussed with the Statistics Sweden management. These diagrams have also played an important part in spreading awareness of the register system within Statistics Sweden, highlighting what deficiencies there are and what important improvements need to be made.

Picture 2.1: Statistics Sweden's register system – Register inventoried in accordance with main type of object



3. The Swedish Basic Registers

The system of basic registers is the foundation of the entire register system and consists of four registers containing important object types. Should one of these basic registers not exist or be of low quality, the entire register system would be considerably less utilisable for statistical purposes. The role of the basic registers is to define the objects in the system, where good unit demarcations and good coverage are crucial with regard to the quality of the system as a whole. The variables which are important in a basic register are those which identify the objects and can be used for linking them with objects in other registers. Indications of time for different events are needed in order that it shall be possible to produce a stock in respect of a specific period or point in time.



Each basic register can be found in four versions:

- Current stock
- Stock at a certain point in time (e.g. 1 January)
- Changes during a certain period (calendar year)
- All objects to be found during a certain period (annual stock)

At present there is no good co-ordination between the basic registers with regard to reference times, but a better co-ordination should be established.

The *Population Register* is to be found in three versions:
- Current stock, updated every 24 hours
- Stock 31/12
- Changes during the calendar year

The *Real Estate Register* represents the situation 1/1 and is updated once a year.

The *Activity Register*, which at present mainly concerns the activity employment, refers to the calendar year and is updated once a year.

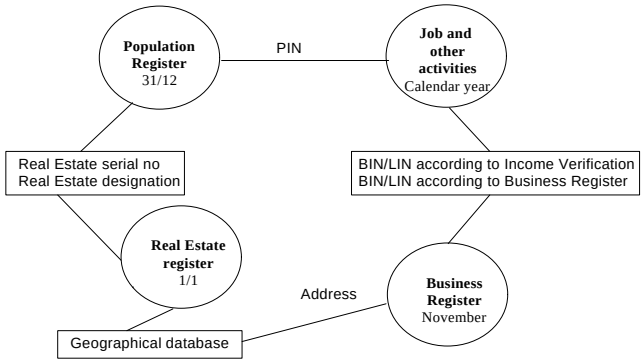
The *Business Register* is updated on a weekly basis and indicates on the one hand the current stock, on the other hand a November version. Annual frameworks for this register have been discussed and have occasionally been produced.

Regarding all these basic registers, work is in progress for effecting a platform change from flat files in a mainframe to databases in NT servers. The largest advance in this respect has been made with regard to *the Population Register.* A common feature is that in the new environment there are being created both an input database, where administrative data and, if applicable, Statistics Sweden data are received, structured and processed, and an output database, used in the production of statistics. The input database has a complicated database structure with many so-called database tables, whilst the output database is simpler and is more like the flat files from the mainframe environment.

4. Linkage between registers

In order that Statistics Sweden's registers shall function as a system it must be possible to link the objects in different registers with the aid *of linkage variables*. An efficient system requires that the linkage variables be of good quality and that the same ones be found in different registers. The links between the basic registers are of particular importance.

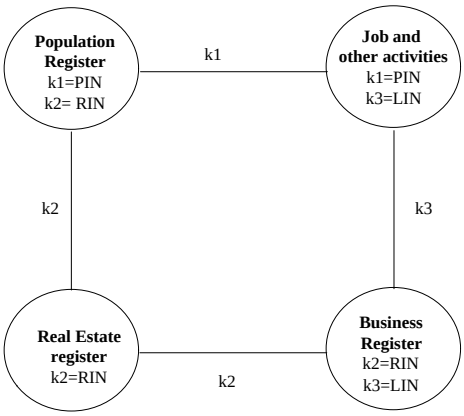
Picture 4.1: Actual linkage between the basic registers[2]



In the present system special linkage registers are needed to link objects in the Activity Register to the Business Register and to link persons in the Population Register to real estate in the Real Estate Register. The reason for this is that neither establishments nor items of real estate have the same identity in different basic registers. In order to link an establishment to the premises where the work is done, there is required a processing with the geographical database. There are certain problems of matching between the Activity Register and the Business Register, and major problems of matching between the latter register and the Real Estate Register

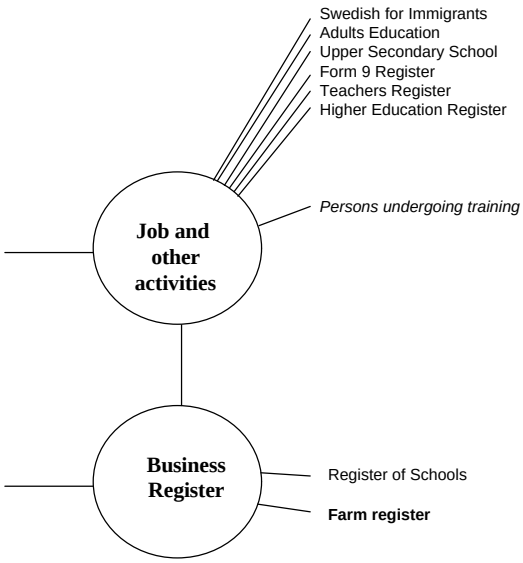
Picture 4.2: Planned linkage between the basic register[3]

[2] LIN – Local activity Identification Number, BIN – Business Identification Number, PIN- Person Identification Number
[3] LIN – Local activity Identification Number, RIN – Real estate Identification Number



Within the framework of the Register Project we are discussing the possibilities of direct links between the basic registers with only three linkage variables (k1-3 in the figure). In the first place the Activity Register and the Business Register would cover the same establishment population with the same identities. Furthermore the same real estate numbers should be found in three basic registers. The gain would be simpler and faster register processing and fewer problems of matching.

Picture 4. 3 Planned linkage between basic registers and statistical registers



Links between the Activity Register and the different registers in the labour market field are established with three variables: personal identity number, establishment number and organisation number. At present there are no links between the Activity Register and the registers on the right.
In the new Business Statistics Register there will be consistency with the industrial statistics population. There should also be consistency with the other three registers on the right

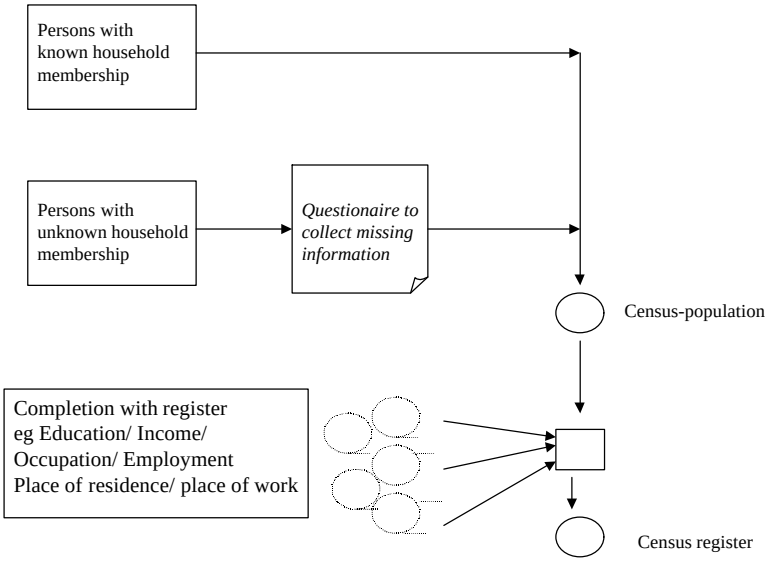
5. Integrated registers make integrated statistics

5.1: The example Census 2000 in Sweden

For a long time the plan was that the census to be carried out in connection with the millennium should be based entirely on administrative registers. However, the Swedish register system has not contained information concerning household membership, information central to the census. Therefore a register of dwellings was to be built up and the population of Sweden was to be registered by dwellings instead of by real estate. This idea, though, has been

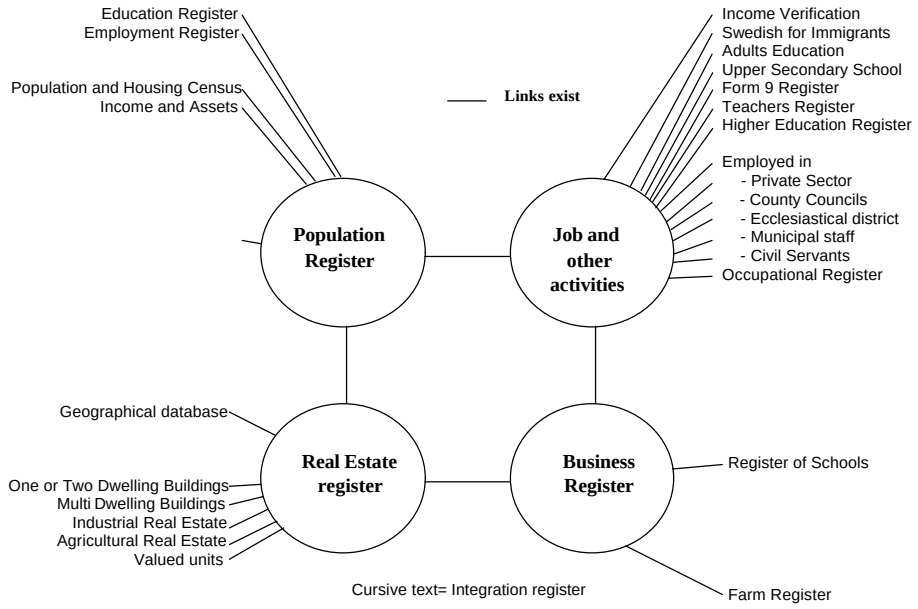
abandoned and at present the plan is to carry out the census by means of registers and questionnaires. Information regarding household membership is therefore to be found by means of a questionnaire when it comes to the part of the population not in single-family housing or having changed address since the last census. The design of the census can be seen from the following picture:

Picture 5.1.1: Overview of sources of information for census 2000



Other information as demography, income, employment status and education will be taken from administrative registers. To be able to use this information in an consistent way the following register and links are to be used

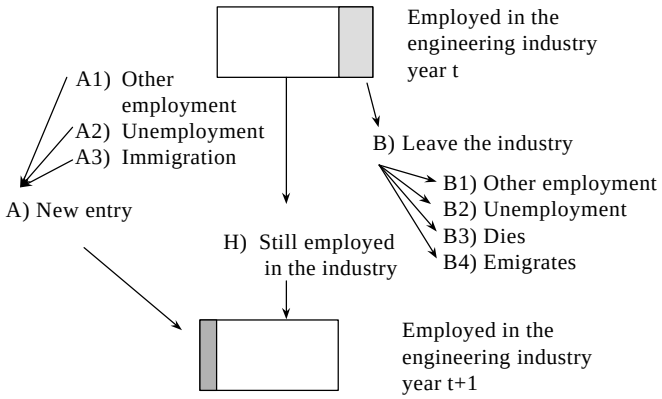
Picture 5.1.2: Register planned to provide Census 2000 with information



5.2: Socio-demographic study of the engineering industry [4]

During the period 1988-92 more than 20% (rather more than 100,000) of those working in the engineering industry left their jobs. There was both union and employer interest in following up what effect this change within the engineering industry had on employment trends and the demand for, and supply of, manpower.
To elucidate these questions it was necessary to link different registers over time. In order to describe the different flows of persons the following rough scheme was used, the basic population consisting of all persons employed in the engineering industry at the end of year t and at the end of year t+1. Through matching these two groups against each other it becomes possible to describe the different flows indicated in picture 5.2.1.

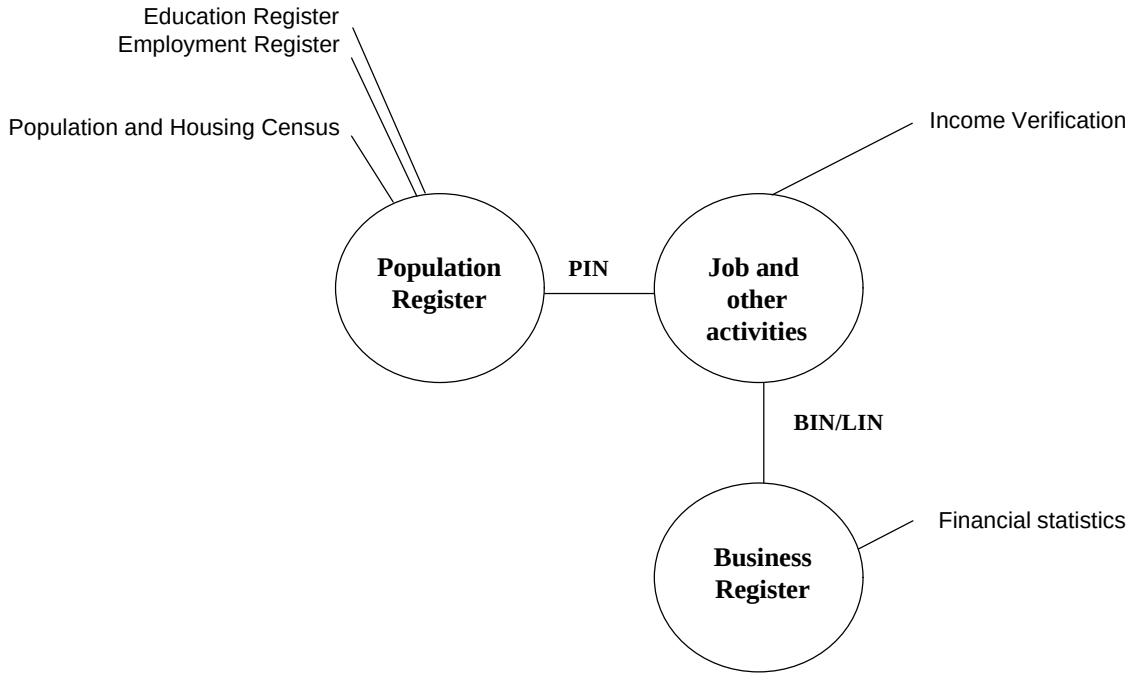
Picture 5.2.1: Flows of staff to and from the engineering industry



In order to define the populations information from the Population Register, the Activity Register and the Business Register must be linked. What this means, roughly, is that information concerning the engineering industry is fetched from the Business Register, and information concerning the persons from the Population Register, whilst the Activity Register links the persons with the companies/establishments. Picture 5.2.2 shows the registers, which were used.

[4] Study by Erik Nyberg, University College of Växjö, on behalf of the Labour Market Board's co-operation group for the engineering industry.

Picture 5.2.2: Registers used, and the links between them[5]



With the aid of these registers and the links between them it is possible to describe each flow in terms of

- Demography (age, sex, socio-economic group, country of birth, occupation)
- Education (highest and most recently acquired certificate)
- Company structure (size of company, location, change in staff volume)
- Business economics (export percentage, change in export percentage, solvency, profit)

The study indicates that the links between the Population Register, the Activity Register and the Business Statistics Register function satisfactorily. On the other hand the links between the financial statistics and the Business Statistics Register were not consistent, which meant that it was impossible to derive a description of the business economics aspects from the register system.

5.3. A longitudinal database for education, income and employment (LOUISE)

In connection with a project studying the exit from the labour market there began a discussion of the more general need for a longitudinal database for describing the social dimensions of the dramatic changes in the Swedish labour market in the 90s. A major project was initiated where information from the labour market field, the social field and the educational field was brought together. The project came to involve a fairly general approach where the main goal has been to produce a database permitting the maximum possible flexibility with regard to describing the labour market from different angles.

5.3.1. Contents

A database has been built up which covers all persons in the 16-64 age group during the period 1990-96. As of 1995, persons older than 64 are also included. The information was collected from several registers where the Population Register gives the population each year. Links between the different registers are provided by Population Identity Number (PIN).

[5] LIN Local unit identity number

During the processing it is possible to identify persons who die or who immigrate or emigrate, and this information is then stored in the database.

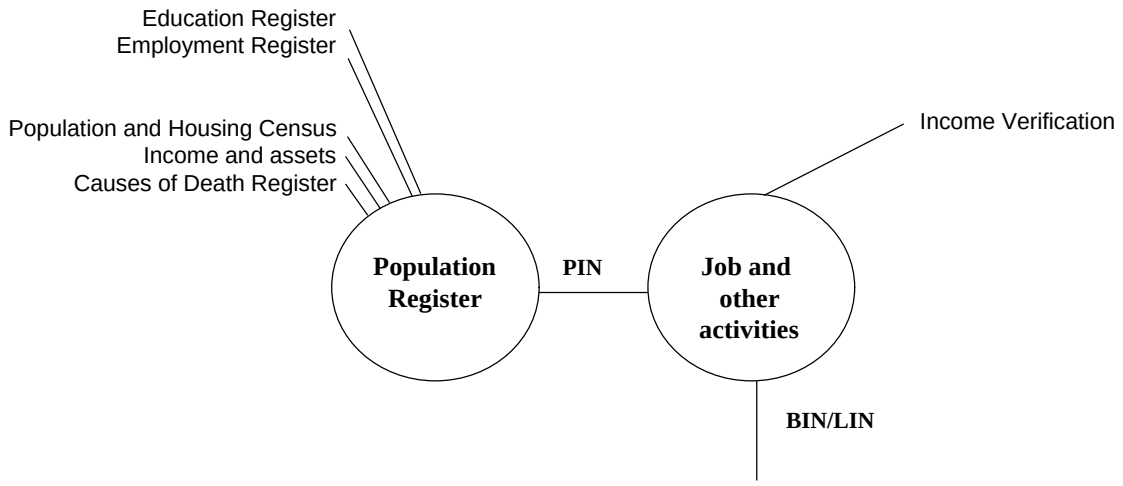
By way of links to the Population Register and the Activity Register it is possible to derive new objects such as household, establishment and company. Variables included in the database are on the one hand pure copies of included register variables, on the other hand (and there are a great number of these) aggregated or derived variables which have been produced. Examples of such variables are

- Income type, where different types of income have been grouped in accordance with a special scheme:
  - Student - National service - Parental leave
  - Ill / Industrial injury - Unemployed - Employment measure
  - Rehabilitation - Partial pension - Disablement pension

- Combinations of social income types related to working-hours. The variable indicates what types of income the person has had during the year.

Every annual section of the database contains some hundred variables stored in ten or so databases tables.

Picture 5.3.1: Registers and links used



5.3.2. Use of the longitudinal database

The database has been used quite extensively in various research projects. Examples of such projects are those concerning

- Labour market changes, economic mobility and welfare
- Mobility on the labour market
- Socio-economic consequences of neurological disease
- Long-term dependence on supplementary benefits

6. Methodological Problems and Preconditions

In order to produce integrated statistics based on registers like the examples presented above, a number of methodological preconditions must be fulfilled. A problem is that - in contrast to the state of affairs regarding traditional surveys - best practice is not always documented and there is more of a basic need for methodological work.

This section is based on a newly started inventory of methodological practice and quality issues in the production of register statistics. To some extent it can be seen as the listing of methodological problems which have to be considered in the future when the development of register statistics is discussed.

Statistics Sweden's sample surveys involve a large number of qualified statisticians and the use of well-documented methods, which have long been discussed in reports and at seminars. Things are different when it comes to register statistics - ad hoc methods are used, these methods are not discussed and have not been well documented, and few statisticians have taken an interest in this field.

In the case of sample surveys the first thing is to determine the population and which parameters are to be estimated for which groups. This in turn determines the character of the survey with regard to sample, data collection, estimation and presentation. Thus there first comes demarcation and definition, then data collection. As a rule one survey at a time is considered, and a limited number of parameters.

In the case of register-based surveys a different approach is taken since data have already been collected and are available in different registers not tailored for a particular application. With the aid of available registers a selection is made of objects and variables relevant to the issue addressed by the survey. It may happen that, on the basis of available registers, new variables - and possibly new objects as well - have to be derived. Thus first there are data, then there is determination of population, parameters and groups. Sample errors do not restrict the possibilities of selecting groups for the coming analysis and report of results. A type of error, which can cause problems, on the other hand, is wrong reporting in the administrative systems. In designing statistical registers and register systems there is a desire to make them flexible and thereby widely applicable. Therefore an extremely important part of the register-statistical methodological work consists of structuring and improving the whole, which is to say finding the best design for the register system. Included in this is the long-term work of watching over and influencing the access to administrative data for statistical purposes.

So even though a large part of the statistical methodology is the same for survey statistics and register statistics, like all non-sampling errors and all problems of analysis and presentation, the ways of thinking are different since the sampling errors and the design problems arising from them are so central to survey statistics.

The register-statistical methodological work comprises the following:

- The work of creating a statistical register with the aid of different administrative sources. This saves costs on data collection and reduces the burden on those furnishing information.

- The work of co-ordinating several statistical registers into an interacting system. This causes existent data to be used more efficiently and increases the chance of producing new statistics.

Register-statistical methodological work is an important part of the process from original provision of information to completed statistical register, as is illustrated here:

Processed to make objects and variables meet statistical requirements (section 6.1):

Editing of basic data
Missing objects and missing values
Linking and matching
Time references
Derived objects
Derived variables
Integration register (section 6.2)
Documentation and Metadata (section 6.3)
Quality assurance (section 6.3)

Used for production of statistics

6.1. Methods of processing

Just as we have made an inventory of registers existing within Statistics Sweden, we also need to make an inventory of methods actually used for the creation of statistical registers. On this subject, however, we are only at the beginning, therefore the discussion of it here will remain rather brief. To some extent, indeed, this section is made up of questions to which we have not yet found good answers.

Editing of basic data

Statistics Sweden has carried out a major review project (SCB 1997). The report presents methods for the review of, in the first place, sample surveys. Macro-review (or selective review), which is the method recommended, is designed for the review of objects and variable values which have a large effect on the estimates in a survey where there are given groups and estimates. This method is less suitable for register statistics, where it needs to be possible to select groups in a variety of ways. An income value, for instance, may appear normal with one division into groups, extreme with another.

Steps taken because of missing objects and measurement values

Here, too, the centre of attention at Statistics Sweden is the problem of non-response in sample surveys. The annual Non Response Barometer indicates only the non-response situation with regard to different sample surveys. Furthermore the non-response manuals which have been produced focus entirely on the problems encountered in sample surveys.

Linking, matching and selecting

The choice of objects and variables in the individual case must be dependent upon considerations of relevance and can only be made upon the basis of a broad and deep knowledge of the origin and contents of (and of any changes in) the different registers which may be used in a matching and linking or selecting. An important prerequisite for the making of a new register is that metadata are available and are from the start planned as a part of the new integrated register. This is particularly important when it comes to the making of longitudinal registers.

The change of platform, which is to say the transition from mainframe to databases in a PC environment which are to be handled with SQL, involves a number of technical problems for Statistics Sweden which have yet to be solved. It is technically difficult, for instance, to link and match registers, which are in different servers. Efforts are being made to find a solution for this. Another thing is that there are no good suitable tools for register processing which can be used in production in the new environment. Specialists can use SQL, but for production there is need of, for instance, a tool for register matching.

Processing of time references

The new Population Register and the new Business Register will contain time information for different register events to a considerably greater extent than before. Furthermore, in the Population Register all events for every person will be saved. This means that the processing of time references will be an extremely important part of register-statistical methodological work in the future. The Real Estate Register could also contain more time references, but for reasons of cost the present register is updated only once a year.

Forming of derived objects

Problems arise when it comes to forming households in the Population Register. With a register of flats we would be able to form dwelling households as derived objects in the same way as is done in Denmark and Finland. If it should prove possible to form, on the basis of available information, derived households at least for certain categories, the cost of future censuses of population and housing could be reduced.
Several important types of derived object will be included in the new Business Register. In the administrative sources there are juridical units which are not always suitable for statistical requirements. For this reason a considerable amount of effort must be expended on forming the objects Business Unit, Establishment, Operational Unit and Local Operational Unit.

Forming of derived variables

With the aid of a number of administrative variables, a statistically meaningful variable is to be formed. There are two possibilities: a rule-based derived variable and a model-based derived variable.

In many situations it is appropriate to form a new variable with a rule, which must be documented and fully warranted. Examples are Age and Disposable Income.

Another method - suitable in more complicated situations - is to use a model, which indicates how the statistical variable depends on the administrative variables. By means of a survey a study is made of how the variables are connected, and a measurement model is created which is then used for forming the derived variable in the register. An example of this is Employment Status in the Employment Register. (See Wallgren & Wallgren 1998.) In order to solve the problem of comparability there was a move in this example from a rule-based method to a model-based one.

6.2. Forming of integration registers

According to the traditional - and still prevalent - way of thinking, the task of forming integration registers is so demanding that it is worth creating permanent registers which are continuously maintained. But there is another way of thinking, whereby the requisite linking and matching is to be performed when a commission comes in. There are two advantages to the latter method: on the one hand it facilitates meeting the needs of the individual client (for completed integration registers cannot be as rich in content as the source registers), on the other hand the problems of integrity diminish inasmuch as one is no longer dealing with large personal registers containing a certain amount of sensitive information.

Owing to technological development and the development of our own register system, the need for permanent integration registers will diminish. With an efficient register system and a well-thought-out organisation of servers there must be every possibility of doing the linking and matching of registers which is required for a particular task and then de-identifying the completed integration register.

What we visualise is that when the register system functions, which means that microdata and metadata are available, those who have the responsibility for a commission or the development of a new product shall easily be able to perform the requisite processing, linking and matching. This brings up important questions of responsibility and organisation. How is the commissioned work to be organised? Who have the authority to publish statistics or carry out commissions, which are partly based on population statistics? And so on ...

6.3. Documentation and metadata

Metadata have an extremely important role when it comes to work on register statistics. In forming integration registers where data from different registers and years are matched and linked, there is great risk of extremely grave error if there is a lack of familiarity with problems of comparability and changes of definition. The work on register statistics requires more metadata than we have at present, and these metadata must be of high quality. All who work with register statistics must contribute to the development of the metadata part of the register system. On the one hand those with responsibility for registers must document, on the other hand all who use the registers must report defects and pass on their views to the person with responsibility for the register in question.

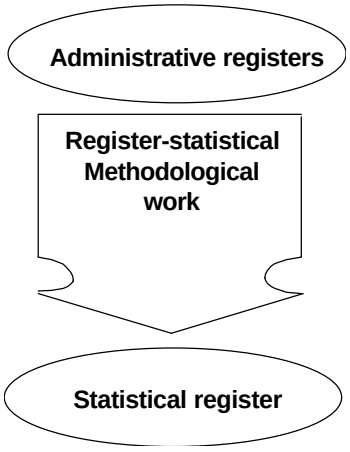
6.4. Quality assurance

The work with the register statistical infrastructure at Statistics Sweden has resulted in an idea about *a general model* for producing register-based statistics in different areas. The model describes the different steps in the production cycle that has to be identified and described for each product. This is a necessary step towards better control and quality in different type of register based statistics. During recent years we have noticed a growing interest in using register as a

source for statistics and information. To link different register as shown in this paper open a lot of possibilities but it will also raise a lot of new demands of quality in different dimension of the register information that we have to be aware of. It is therefore important to have tools and methods to be able to help the user of register based statistics with quality information about strong and weak part in the information he or she will receive.

The idea of a quality model for register based statistics is briefly described in picture 6.2.

Picture 6.2: Brief description of a quality model for register based statistics



Quality assurance

Supplier contacts
Reception control
Cause and extent of non-response, imputing
Cause and extent of non-matching objects
Evaluate quality of objects and variables
Inconsistencies between registers are documented and reported
Defects in metadata are reported

Defects are reported back to the level close to the source of the error

References

Statistics Denmark and Eurostat (1995): Statistics on Persons in Denmark. *A register-based statistical system*.

Carlsson, F (1996) *"Use of Labour Force Survey(LFS) to improve the quality of statistics,based on administrative registers*" Proceedings of IAOS 5th Independent conference, Reykjavik 2-5 July 1996

Eurostat (1997): *Proceedings of the seminar on the use of Administrative Sources for Statistical Purposes*, Luxembourg, 15-16 January 1997.

SCB (1997): *Granska effektivt!* Rapport från granskningsgruppen mars 1997.

Sandgren, Per, Wallgren, A. and Wallgren, B. (1994*) "Measuring Employment using Income Verifications from Tax Records"* Unpublished work

Selander, R., Svensson, J., Wallgren, A. and Wallgren, B. (1998): *Administrative Registers in an Efficient Statistical System - New Possibilities for Agricultural Statistics? HOW SHOULD WE USE IACS DATA?* Eurostat and SCB 1998.

Wallgren, A. and Wallgren, B. (1997): *The Role of Administrative Registers in an Efficient Statistical System - Methodological Problems and Quality Issues*. Eurostat: Proceedings of the seminar on the use of Administrative Sources for Statistical Purposes, Luxembourg, 15-16 January 1997.

Wallgren, A. and Wallgren, B. (1998): *Svensk Registerstatistik – Nuläge och Utvecklingsmöjligheter*. Contribution to the Nordic Statistics Meeting in Lillehammer, Norway, August 8, 1998.