To understand the Possibilities of Administrative Data you must change your Statistical Paradigm!

Anders Wallgren and Britt Wallgren
Örebro University and Statistics Sweden, ba.statistik@telia.com

Abstract

This session's title is: *Is New Emphasis on Administrative Data a Supplement to Survey Research or a 'Paradigm Shift'?*

To be able to answer this question we discuss the development in the Nordic countries in Europe where administrative data have been used for statistical purposes during about 50 years. We try to explain how the production systems at the Nordic statistical agencies differ from the production systems in countries that do not use administrative data to that extent. When many registers can be linked, the register system itself is a new important factor that has great impact on the way statistics is produced. We also compare record linkage in different countries and discuss the nature of administrative data.

Key Words: Administrative data, register system, record linkage, register-based census, register-based national accounts.

1. What can be done at NSOs with full access to administrative data?

The National Statistical Offices (NSOs) in Norway, Sweden, Finland and Denmark have today access to large amounts of administrative data that are used for statistical purposes. The preconditions for using administrative data in this way have been very good in the Nordic countries:

- Good civil registry, in Sweden's and Finland's case at least since 1748. All persons have been registered at the property where they live and births, deaths, marriages and migrations have been recorded since then.
- Unique national personal identity numbers were introduced early, in Sweden's case during 1947. These identity numbers have been used in all administrative systems in the public sector for a long time and the same numbers are also used by banks, insurance companies etc.
- When computers were introduced in public administration, the earlier paper form registers were replaced by data files and during the 1960's the NSOs started to use these administrative registers. This is described in Unece (2007).
- More and more statistical registers were gradually created and today the NSOs in the Nordic countries have access to large amounts of administrative data that has been used to create systems of statistical registers were many registers can be linked with identity numbers and address codes.

These systems of statistical registers have been used to replace the traditional questionnaire censuses in these countries by register-based censuses during the period 1981 - 2011. It is today also possible to use economic administrative data to create partially register-based national accounts.

Statistical registers are used for the production of official statistics and as a basis when integrated registers for research are created. Some of these registers for medical and social science research are longitudinal and cover about 25 years.

2. The transition to a production system with many registers

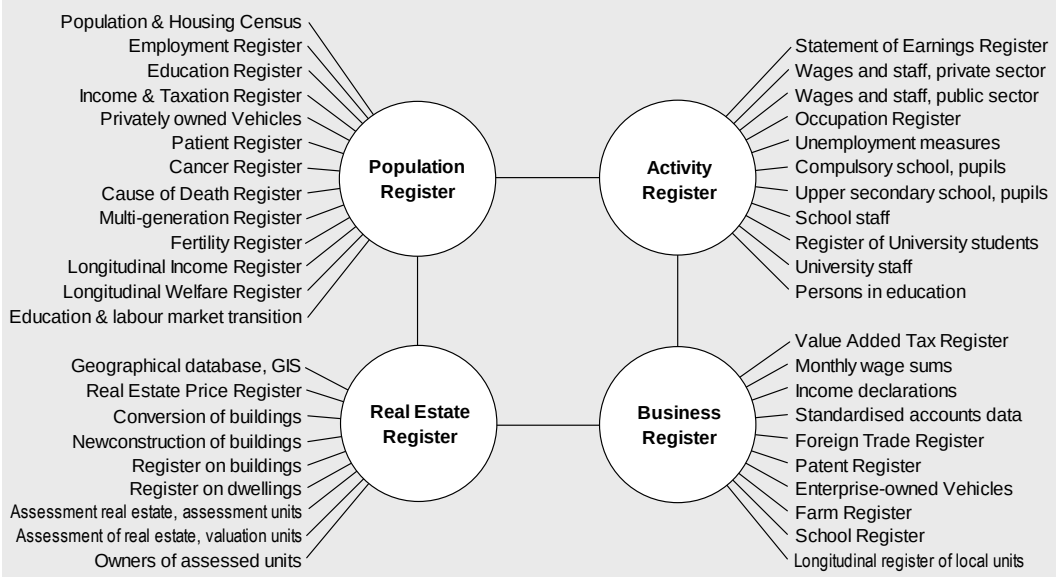
Before the 1960's no registers were used at Statistics Sweden. Today we have about 50 statistical registers at Statistics Sweden and about 70 statistical registers for different medical diagnoses at the National Board of Health and Welfare.

Instead of sampling methods based on maps or address lists, all sampling today in the Nordic countries is based on registers:

- All frames are created with registers, mainly the Population Register and the Business Register. Only one-stage sampling designs are used.
- Our interviewers never ask about identity number, sex, age, income or education etc. Our questionnaires to persons are also free from questions about these issues. Enterprise surveys also use economic activity and institutional sector from the Business Register.
- We use many register variables (e.g. about 50) for stratification and as auxiliary variables during estimation and for calibration to adjust for nonresponse.

Today, all our surveys use our statistical registers – a great part of our production of statistics is based entirely on our registers and all sample surveys also are dependent of our system of statistical registers that is described in Chart 1 below.

Chart 1: A system of statistical registers ordered by object type in four categories – persons/households, activities (jobs/studies), enterprises/organizations and real estate



3. What is a register system?

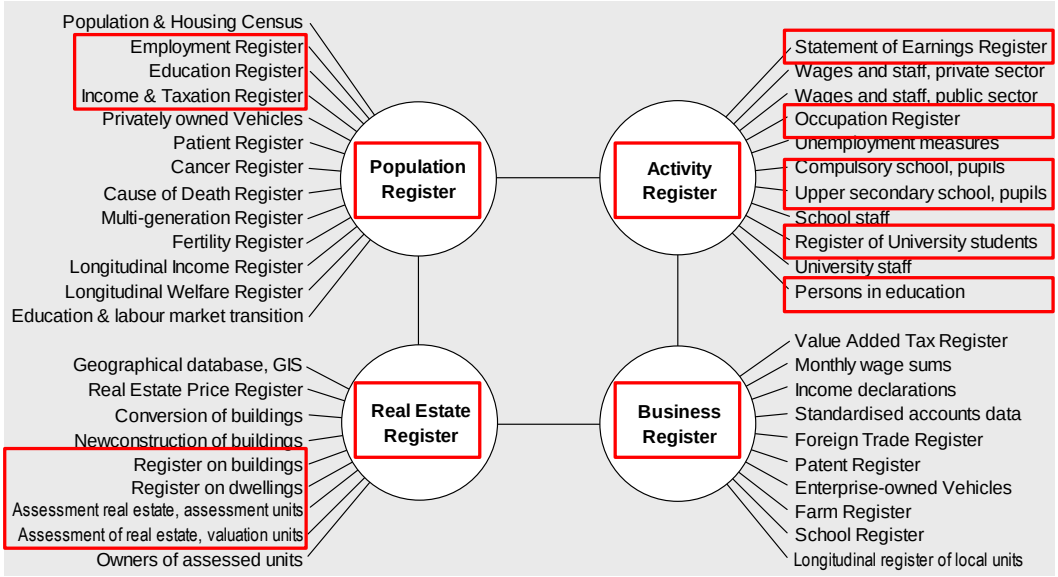
Up to now we have explained that all Nordic countries have public administrations that generate administrative data based on identity numbers of good quality and that the NSOs in these countries have access to all this data for statistical purposes. This right to use administrative data is combined with laws and rules that protect privacy and confidentiality of citizens and enterprises.

However, the register system itself is another factor that opens new possibilities and requires new statistical methodology. In countries that have started to use administrative data for statistics it is natural to use *one* source to create *one* register that is used to produce *one* kind of statistics. One example of this is that *income declarations* from persons are used to create an *income register* and that this income register is used to produce *income statistics*.

But when a NSO has got a sufficient number of registers that can be linked, this system of registers itself can be the source of new register-based surveys. If a new demand for information arises then by making a new combination of already existing sources in the system it will in many cases be possible to answer the new demand for statistical information. In Chart 1 above there are a number of important registers that have been created in this way: the Employment, Education, Multi-generation, Longitudinal Income, Longitudinal Welfare and Labor Market Transition registers are of this kind.

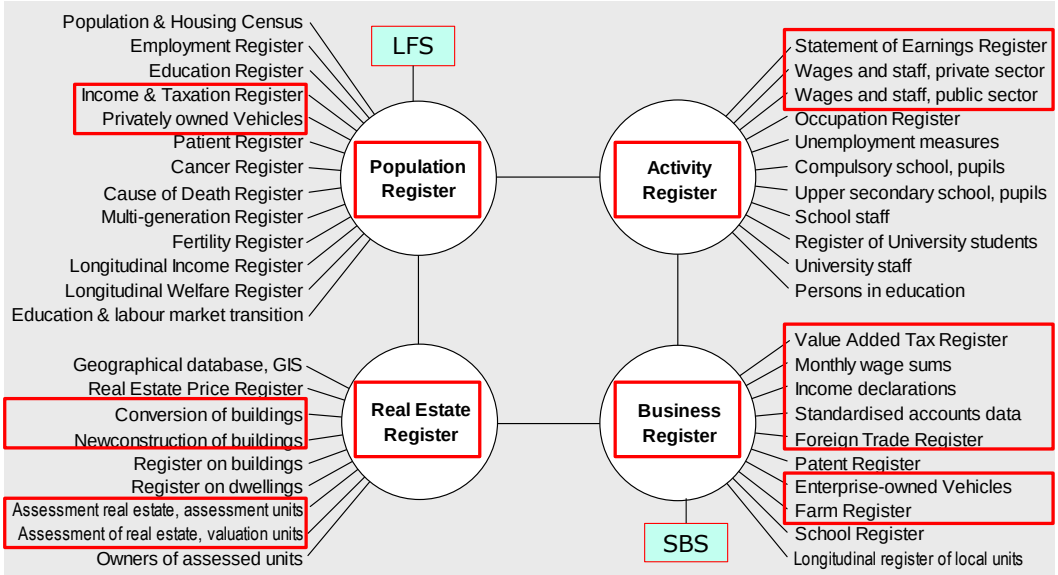
Also the register-based Population and Housing Census is created by making a new combination of data in already existing statistical registers in the system. This is illustrated in Chart 2 below where the registers used for the census are marked red. The main reason for the register-based census is to save costs but also the quality can be better as nonresponse will not be a problem. The long form in a tradition census is made unnecessary by combining registers from different subject matter areas such as employment, education and income etc.

Chart 2: The register-based census is based on a large number of registers in the system



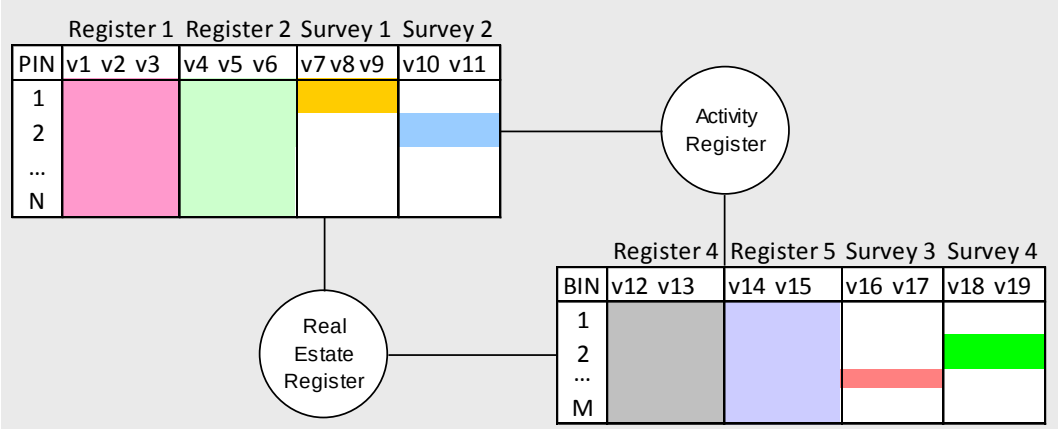
Also sample surveys and statistical registers can be combined. One example here is that enterprise data, both from sample surveys and administrative registers, can be integrated to produce consistent and coherent estimates for the National Accounts. In Chart 3 below this is illustrated with registers marked red and two sample surveys, the Labor Force Survey (LFS) and the Structural Business Statistics Survey (SBS) marked turquoise. The main reason for this combination of data is to find and reduce errors.

Chart 3: Sources used to produce consistent and coherent estimates for National Accounts



The thin lines in the charts above show how the statistical units in registers and sample surveys can be linked. Only three linkage variables are used in a system of the Nordic kind – Personal Identification Numbers (PIN), Business Identification Numbers (BIN) and address codes. All registers and sample surveys on persons can be combined or integrated into one data matrix by Personal Identity Numbers. In a corresponding way, all registers and sample surveys on enterprises can be combined or integrated into one data matrix by Business Identity Numbers. In Chart 4 these two combined data matrixes are illustrated. Register 1-2 combine micro data for the whole population, (sample) surveys 1-2 can be combined with register 1-2 for the persons in each sample. In a corresponding way business data from register 4-5 and sample surveys 3-4 can be combined.

Chart 4: Integrated data matrixes for persons and enterprises. v1 – v19 denote variables



The existence of a register system of this kind has the following consequences:

- All registers and censuses in the system can be combined. Not only all registers on persons can be combined, but also all registers on persons can be combined with all registers on enterprises. E.g. if the Employment Register is combined with the Foreign Trade Register via the Activity Register, enterprises with foreign trade can be described by the proportion of staff that has university education.
- All sample surveys can be combined with all registers. E.g. if the Labor Force Survey (LFS) is combined with the Foreign Trade Register via the Activity Register, each employed person in the LFS can be classified as working in an exporting enterprise or not.
- No sample surveys can be combined with any other sample surveys as different parts of the frame populations are included in different sample surveys. This means that at a NSO with no registers, it is quite sufficient to think of *one survey at a time*. "One survey at a time"-thinking dominates the sampling theory and also the literature on (sample) survey methodology and quality. This way of thinking was adequate before the register system existed, but at a NSO with a register system this old thinking or paradigm leads to that the potential of the register system based on administrative data is not fully used.

Combining sources has two important statistical advantages:

- From subject matter point of view the combined data matrix has a much richer content.
- From methodological point of view the combination of different sources makes it possible to detect and correct errors that are not seen when we analyze only one source at a time.

4. Record linkage in the Nordic countries

In the Nordic countries record linkage is done by exact deterministic matching[1] of records in different sources. As a rule only one linkage variable at a time is used when two sources are combined – an identity number or an address code. This method of record linkage is used to combine up to 125 population registers in Statistics Sweden's largest longitudinal register that is used by researchers. The quality of this large scale record linkage is high.

We would describe this linkage method as a *register system approach*. Of course, not all identity numbers used in Swedish administrative sources are correct, but to deal with this quality issue the work is organized in a specific way:

- Special units at Statistics Sweden are responsible for one of the *base registers*[2] and the base register is used to create one or more *standardized populations* that are used by other units that work with register-based statistics.
- The unit that is responsible for e.g. the Population Register is also responsible for the PIN-variable. This means that they must keep track of persons that are

[1] We use here the terms given in Herzog, Scheuren and Winkler (2007)
[2] The role of base registers and standardized populations is described in Wallgren and Wallgren (2007)

allowed to change their PIN or replace a preliminary PIN with a definitive PIN. Old and new PIN is included in a cross reference table together with the date when the change occurred. Sometimes an immigrant gets a PIN from a dead person. Also such information is recorded by those who work with the Population Register. With this information the staff at the Population Register can edit administrative data on persons, replacing the PINs in the source with corrected PINs from the Population Register.

- When the standardized populations are created that are used by other register units, persons with incorrect or unknown PIN are excluded. Such persons can be foreigners studying at Swedish universities and these will be excluded from the population when only persons permanently living in Sweden are included.
- The link between the Business Register and the Real Estate Register is the link that is problematic in the Swedish system. We have only mailing address and due to that spelling of street names is not standardized and some enterprises have post office box we have problems with mismatch.

The record linking methods developed by Felligi and Sunter (1969) are not used in the Nordic countries. Their exact but probabilistic matching method was developed to deal with situations where you have no high quality identity number that is used in many sources. In such situations it is necessary to combine a number of linking variables such as name, address, birth date and birthplace etc.

If we compare record linkage in the Nordic countries and e.g. the U.S. quite different methods are used today. But for us it seems reasonable that the social security number in the future can be used in the U.S. in a similar way as we use PIN in the Nordic countries today. If you want to match e.g. 125 population registers in the U.S. in the future, then exact deterministic matching is the best solution.

5. The nature of administrative data

Data that has been collected or created by administrative authorities can be of different nature. Some data are actually statistical data, if the authority wants to produce its own statistics. E.g. the Swedish Public Employment Service produces its own statistics on jobseekers and some variables collected from the jobseekers can be statistical data.

Other kinds of variables are legally important – if you give wrong information on these then you have done something illegal and can be punished. E.g. the information taxpayers give in tax forms is of this kind. In Chart 5 three kinds of data are compared. The statistical questionnaire (1) and the tax form (2) may look similar, but the cognitive processes are different. In (1) you can answer what you want and if you want, but in (2) it is your *duty* to understand the questions, you *must* retrieve the information you need, your judgment *must* be according to the law and you *must* report before a certain date.

Chart 5: Statistical questionnaire data, tax form data and administrative decision data

1. Questionnaire from NSO:	
Turnover	100
Costs	90
Profit	10

2. Tax form:	
Turnover	100
Costs	90
Profit	10

3. Decision by Tax Board:	
Turnover	100
Costs	70
Profit	30
Tax	9

A third category of variables are those that are decisions made by the authority (3). The Tax Board decides on taxable income and the amount of tax that should be paid, a court decides that a person is guilty of crime against a certain law and should get a specific punishment, social authorities decide that you should get some kind of benefit and how much money you will get etc.

We can take the administration within a manufacturing enterprise as an example of administrative data that is purely administrative in nature:

Chart 6: Administrative data based on decisions, without data collection or measurement

A customer phones and asks if enterprise X can deliver a certain quantity of a certain commodity. How much will it cost and when can it be delivered? After negotiations the following administrative data has been created:

Customer identity:	xxxx
Article number:	yyyy
Quantity:	qqqq
Price:	pppp
Delivery date:	dddd

This kind of administrative data can afterwards be used for a register-based survey on sales. It should be observed that here there is no measurement and no collection of data – data is generated by the administrative process. A statistical measurement is of a quite different nature: Then the true values of the variables exist first, and then we measure and collect the data.

We have many times heard that administrative data have the same character as statistical data, so all our theory on measurement errors etc. can be applied. "No new quality issues exist; our books on survey quality (*sample* survey quality according to us) have all quality issues that are important for administrative data". We disagree to this. In the enterprise example above there is no measurement and thus there are no measurement errors.

6. The new paradigm

This session's title is: *Is New Emphasis on Administrative Data a Supplement to Survey Research or a 'Paradigm Shift'?*

Our paradigm determines how we perceive statistical surveys and administrative data. Your statistical paradigm is your way of thinking about statistical issues, and most persons are not aware of her/his paradigm. Before the discussion on how to use administrative data for statistical purposes started, the statistical paradigm was based on probability theory and inference theory. This paradigm dominated at statistical departments at universities and also among methodologists at statistical agencies. Sampling theory is defined within this paradigm.

In Wikipedia we find the following text:

The historian of science Thomas Kuhn gave paradigm its contemporary meaning when he adopted the word to refer to the set of practices that define a scientific discipline at any particular period of time. Kuhn himself came to prefer the terms exemplar and normal science, which have more

precise philosophical meanings. However in his book The Structure of Scientific Revolutions Kuhn defines a scientific paradigm as:

- what is to be observed and scrutinized
- the kind of questions that are supposed to be asked and probed for answers in relation to this subject
- how these questions are to be structured
- how the results of scientific investigations should be interpreted

When we started our work with register statistics, we found that no methodologists at Statistic Sweden had worked with this before us. In the other Nordic countries it was the same – register statistics was created by subject matter people and IT-staff only. Some methodologists said to us: "Register statistics is a census and a census is as a sample survey with $n = N$."

So, according to the dominating statistical paradigm at that time, there was nothing of interest to statistical science to be observed and scrutinized regarding register statistics and there were no important statistical science questions that could be asked regarding statistical registers. The real fact is that the Nordic register-based statistics since the 80's was created outside the realm of statistical science, with the consequences that ad hoc methods have been used instead of statistical science methods. If we don't like the present ad hoc state, then a radical change is needed. So our answer to question above is a clear *"Yes, the new emphasis on administrative data requires a paradigm shift!*

This new paradigm cannot be based on probability and inference theory as these historically important parts of statistical science can't help us when we work with registers. Instead micro integration of many sources must build on a theory of statistical systems or a theory for the whole production system at a NSO. Sampling theory is oriented towards thinking on one survey at a time; we cannot think in that way when we work with administrative data and try to fit different sources into the register system.

What should be the meaning of the term "survey"? This is also an important paradigm issue. We think that Statistics Canada's way of defining the term survey is the best; they distinguish between the following four kinds of surveys:

Chart 7: Statistics Canada, Quality Guidelines, Fifth Edition – October 2009

The term survey is used generically to cover any activity that collects or acquires statistical data. Included are:

- a census, which attempts to collect data from all members of a population;
- a sample survey, in which data are collected from a (usually random) sample of population members;
- collection of data from administrative records, in which data are derived from records originally kept for non-statistical purposes;
- a derived statistical activity, in which data are estimated, modeled, or otherwise derived from existing statistical data sources.

The guidelines are written with censuses and sample surveys as the main focus. Very often the term survey is used as a synonym for sample survey, as ASA does. We think that the term survey is more general – there are four different kinds of survey that use sometimes different methods but have the same aim: To collect or acquire data to be able to produce estimates. If we decide to produce estimates of e.g. employed by industry with

a census, with a sample survey or with a register survey, then we do a survey in all these three cases. However, it is a choice between different survey methodologies.

References

Fellegi, I. P., Sunter, A. B. (1969): A theory for record linkage. *Journal of the American Statistical Association*, 64, pp. 1183-1210.

Herzog, T., Scheuren, F., Winkler, W. (2007): *Data Quality and Record Linkage Techniques*. Springer.

Unece (2007): *Register-based statistics in the Nordic countries – Review of the best practices with focus on population and social statistics.* United Nations Publications.

Wallgren, A., Wallgren, B. (2007): *Register-based Statistics – Administrative Data for Statistical Purposes.* John Wiley & Sons Ltd.