Research and Development – Methodology reports from Statistics Sweden 2011:2

Quality Assessment of Administrative Data

Statistiska centralbyrån Statistics Sweden

Thomas Laitila
Anders Wallgren
Britt Wallgren

The series entitled "**Research and Development** – **Methodology Reports from Statistics Sweden**" presents results from research activities within Statistics Sweden. The focus of the series is on development of methods and techniques for statistics production. Contributions from all departments of Statistics Sweden are published and papers can deal with a wide variety of methodological issues.

Previous publication:

2006:1 Quantifying the quality of macroeconomic variables

2006:2 Stochastic population projections for Sweden

2007:1 Jämförelse av röjanderiskmått för tabeller

2007:2 Assessing auxiliary vectors for control of nonresponse bias
in the calibration estimator.

2007:3 Kartläggning av felkällor för bättre aktualitet

2008:1 Optimalt antal kontaktförsök i en telefonundersökning

2009:1 Design for estimation: Identifying auxiliary vectors to reduce
nonresponse bias

2009:2 Demographic methods for the statistical office

2011:1 Three Factors to Signal Nonresponse Bias – With applications to
Categorical Auxiliary Variables

.

Research and Development – Methodology reports from Statistics Sweden 2011:2

Quality Assessment of Administrative Data

Thomas Laitila
Anders Wallgren
Britt Wallgren

Statistiska centralbyrån
2011

Research and Development – Methodology reports from Statistics Sweden 2011:2

Quality Assessment of Administrative Data

Statistics Sweden
2011

Producer Statistics Sweden, Research and Development Department
SE-701 89 ÖREBRO
+ 46 19 17 60 00

Enquiries Thomas Laitila, +46 19 17 62 18
thomas.laitila@scb.se

Source: Statistics Sweden, Research and Development – Methodology Reports from Statistics Sweden, *Quality Assessment of Administrative Data.*

Cover Ateljén, SCB

ISSN 1653-7149 (online)
URN:NBN:SE:SCB-2011-X103BR1102_pdf (pdf)

This publication is only published electronically on Statistics Sweden's website www.scb.se

Preface

The authors work within the European Commission's research project *BLUE-Enterprise and Trade Statistics, Work Package 4: Improve the use of administrative sources,* FP7-SSH-2009-A, Grant Agreement Number SSH-CT-2010-244767. This report has been prepared as a part of that project.

Statistics Sweden, March 2011

Lilli Japec

Disclaimer

The series Research and Development – Methodology reports from Statistics Sweden is published by Statistics Sweden and includes results on development work concerning methods and techniques for statistics production. Contents and conclusions in these reports are those of the authors.

Contents

1 Administrative Data in Survey Design

The aim of this paper is to discuss methods that can be used when an administrative source is evaluated from a statistical point of view. Should it be included in the production system of a National Statistical Institute (NSI)? How should it be used within the production system? The statistical usability of the source is analysed with a method that uses a number of quality indicators.

The methods discussed in this paper are general, but we have chosen to primarily discuss enterprise and trade statistics.

1.1 Survey Design

Statistical surveys are traditionally associated with *sample surveys*, where part of the population is selected at random and estimates are derived from sample responses. Dalenius (1969) discusses the problem of selecting the design of a sample survey with respect to the objective of the sample survey. Several text books on survey sampling are available, e.g. Cochran (1977) and Särndal, Swensson and Wretman (1992).

There are however several other possible sources for collection of data and information generation. Today there is an increasing interest among statistical agencies in using administrative data for production of official statistics (Wallgren and Wallgren, 2007).

Adding administrative data sources, the researcher has the additional option to choose a *register survey*, a survey based on data from administrative sources. Laitila and Holmberg (2010) deals with the comparison of sample and register surveys in terms of mean squared error focusing on the trade-off between relevance and precision, a major issue in comparison of register and sample survey designs (Holt, 2001).

The researcher also has the option of using a combination of sample and register surveys. Certainly sample surveys and register surveys can be designed in different ways so the option today in survey design is to choose among a set of sample surveys, register surveys and combinations of sample and register surveys.

In traditional sample surveys the researcher is in control of the data collection and can, in addition to sampling errors, evaluate the quality of statistics produced.

The situation is different using register surveys since the researcher has no control over the data collection and registration phase. Some quality aspects can be evaluated using information on the register, e.g. relevance aspects such as population and variable definitions, while other aspects are more difficult to evaluate, e.g. measurement errors. However, a survey design is built up by different components and administrative registers may be included in a survey for different purposes.

1.2 Statistical Quality of an Administrative Source

An administrative data source may be used for other statistical purposes than being the primary source of data for statistics (e.g. Eurostat, 2003). The quality requirements on an administrative register therefore depend on the role of the register in the survey design.

One major use of registers is the formation of a frame for sample surveys. For this purpose the major quality requirements are coverage, possibility to identify population units, contact information, domain indicators and information on auxiliary variables. Using an administrative register or source for a pure register survey puts the same requirements on the register as if it was to be used as a frame (except contact information and auxiliary variables); with the addition it must include data on the study variables. These must also be with small errors.

Even though some administrative registers fail to meet the requirements for high quality of register statistics they can sometimes be combined, adjusting for the shortcomings of the separate registers, as illustrated by Wallgren and Wallgren (2007).

Furthermore, one could think of situations where an administrative source contain insufficient information on primary keys whereby it is not possible to link the units in the source to units in the base register[1]. Such a source can still be useful if the population for the source is well defined and with small coverage errors. One

[1] The base registers define the populations in the Production System (Chart 7 in section 4.3)

application could be to use it for evaluation of register statistics derived from other administrative sources.

In other cases an administrative register may provide important auxiliary information for sampling design and/or estimation purposes even if the register variables are not relevant enough for direct statistical purposes.

The discussion above highlights the problem of establishing general statistical quality of an administrative register. The quality of a register is established with relation to a specific intended use. As a register can be used in more than one way, the quality requirements on an administrative register therefore depend on the overall role of the register in survey design within the production system of the statistical agency.

Thus, a general and useful system for quality assessment of administrative registers can not originate from *one* specific application and the quality of statistics derived in that application. Quality assessment of a register should instead focus on available information on the administrative register and on information that is based on a systematic analysis of the administrative source.

An outline of this kind of analysis is given in this paper. This analysis will enable researchers to judge the quality of derived statistics when the register is used as a component in survey design in all intended applications.

1.3 The Statistical Production System

It becomes natural to think about administrative registers as inputs in a production system, i.e. inputs to a production function. Raw material can in general not be directly used in the production process; it has to be prepared, e.g. cleaning of recycled fibre in paper mills. Substitute raw materials may imply a difference in the quality of the final product and the efficiency of the production process, e.g. virgin fibre gives stronger paper than do recycled fibre. Also, the production technique used may not be defined for some inputs.

The simple analogue to the paper mill example illustrates that the quality of administrative data has to be looked upon from two different views, from the view of the consumer of statistics and from the view of the producer of statistics.

The consumer view concerns the quality of the final product, or the *"Output quality"*.

The producer view concerns two problems:

i) *"Input data quality"* – the preparations of the input needed for use in the production process and,

ii) *"Production process quality"* – the gains in production efficiency of using the input.

To develop a quality assessment system of register data, each one of the three concepts output quality, input data quality and production process quality must be divided into a set of components describing different aspects on the quality concepts.

1.4 Overview

This paper presents a suggestion of indicators to assess the quality of administrative data sources. This quality assessment is performed when an administrative source is discussed within a statistical agency – should we use this source, and how? We also describe the work process where these indicators are derived and analysed.

The suggested indicators are derived from the above division of quality into the concepts of output, input data and production process quality. Components and indicators of the three quality concepts are considered in Section 2. Section 3 provides with a comparison and summary of the suggested indicators. A discussion of results and further problems to be addressed are saved for the final section.

Wallgren and Wallgren (2007, Chapter 10.1) mention that the quality of a statistical register in the Production System depends on three factors (we have adjusted the wordings to fit the present report):

- the administrative sources on which the register is based which determine output quality, input data quality and production process quality,
- the possibilities offered by the Production System to improve the input data quality of these sources which also determine the input data quality of the new source and
- the processing done or methods used to produce the register.

Daas et al. (2008, 2009, 2010) discuss methods to evaluate the statistical usability (or quality) of administrative sources. A number of indicators are developed and are used to analyse each source. On the basis of these indicators the NSI should decide to use or not to use a specific source.

2 Three ways of using a Source

An administrative source can be used in three different ways in the production system of a NSI. In Chart 1 these three ways are illustrated.

1. If the statistical quality is sufficient, the source can be used almost as it is for a statistical product (SP). To see if the source can be used in this way its quality should be judged according to the traditional indicators in Chart 2 below that are used by the NSI for measuring *Output Quality*.
2. We distinguish between two cases:

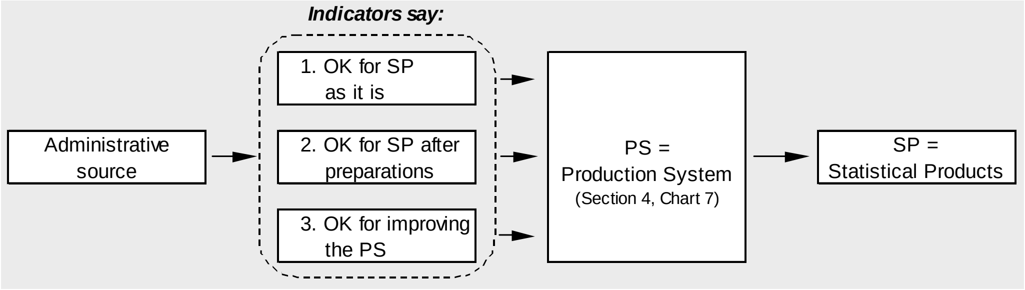
   a) There are some quality problems with the source, but after preparations where other sources are used it is possible to use the source for a statistical product (SP).

   b) The source is already used for some statistical product, but after special register-statistical processing where the source is combined with other sources it is possible to create a new more advanced statistical product (SP).

   When a source is used in one of these two ways, its *Input Data Quality* should be judged.
3. The source can be used to improve the Production System (PS) at the NSI. In this way the quality of existing sample surveys and/or register surveys can be improved. When a source is used in this way, its *Production Process Quality* should be judged.

It should be noted that the same administrative register or source can be used in different ways and therefore it is quite possible that all three quality concepts can be used when a specific source is analysed.

Chart 1. Three ways of using an administrative source in the Production System



2.1 Output Quality

The strongest requirements on an administrative register are found when it would be used as the single source for producing statistics. It is suggested to use such an intended application together with a standard system of quality declaration of statistics as a starting point for deciding what register indicators are important to measure in order to establish output quality.

Chart 2. Indicators for establishing output quality of administrative registers. Statistics Sweden's model for quality declarations.

Quality component	Quality subcomponent	Quality indicators
Relevance	Population and units	Population definition Definition of units
	Variables	Definition of variables
	Reference time	Reference time
	Study domains	Domain variables
	Statistical measures	*(No indicator measured)*
	Comprehensiveness	Scope of content
Accuracy	Frame coverage	Under coverage Over coverage
	Non-response	Missing values
	Measurement	Measurement errors
	Sampling	*(No indicator measured)*
	Data processing	*(No indicator measured)*
	Model assumptions	*(No indicator measured)*
	Overall accuracy	*(No indicator measured)*
	Accuracy measures	*(No indicator measured)*
Timeliness	Frequency	Update frequency
	Production time	Time for data deliverance
	Punctuality	Uncertainty of data deliverance time
Comparability/ Coherence	Comparability between domains	*(No indicator measured)*
	Comparability over time	Comparability over time
	Coherence with other statistics	*(No indicator measured)*
Availability/Clarity	Dissemination forms	*(No indicator measured)*
	Presentation	*(No indicator measured)*
	Documentation	*(No indicator measured)*
	Access to micro data	*(No indicator measured)*
	Information services	*(No indicator measured)*

There are several different quality aspects on official statistics. Following the quality definition in Statistics Sweden's quality model, output quality is divided into the components *relevance, accuracy, timeliness, comparability/coherence,* and *availability/clarity*. The components and their subcomponents are shown in Chart 2 below.

Each component contains one or several subcomponents. For instance, the component comparability and coherence is divided into:

i) comparability over time,

ii) comparability over domains and

iii) coherence with other statistics.

Considering these subcomponents it is realized that some of them contain relative information, constituting a comparison of quality aspects among statistics. E.g. the subcomponent coherence with other statistics involves a comparison of populations, variable definitions and reference times among published statistics from different surveys. Measurements of such quality subcomponents are covered by measurements of other subcomponents; new indicators are not needed.

Also, some subcomponents concern aspects not directly related to the data source, e.g. model assumption in the accuracy component in Chart 2. Such subcomponents are not to be measured via indicators on the register used.

Chart 2 presents suggested quality indicators to be measured on the administrative register in order to be able to assess output quality according to the Statistics Sweden's model for quality declaration of statistics that corresponds to the ESS model.

Notice that here we are only concerned with what should be measured on the administrative register in order to be able to document the quality of statistics produced *using the register as the single data source*. Out of a total of 25 quality subcomponents, 12 are to be measured via indicators measured on the register. There are 13 indicators that are not measured when an administrative source is analysed in this context.

2.2 Input Data Quality

Input data quality concerns the possibility to use the administrative register received by the NSI and the necessary preparations needed for inclusion of the register in the statistical production system. Even if the administrative source *can't be used as it is* for a Statistical Product (SP), because the output quality is not sufficient, it may be possible that the source can be added to the Production System (PS) after some preparations. Here it becomes natural to evaluate input data quality with respect to a base register. As we here discuss enterprise and trade statistics, the base register to be considered is the Business Register.

First it may be necessary to solve some technical problems – the data must be delivered in an appropriate format, this issue should be solved in cooperation with the administrative authority that is responsible for the source; today, this issue should not raise any difficulties and will not be discussed here. Also the linking variables may need to be transformed so that it has the format that is used within the Production System.

A more difficult issue arises if the object type of the units in the administrative source differs from the object types in the base register we want to link to the source in question. The administrative units in the source can then be aggregated into units that can be linked with the base register or a cross-reference register can be created with links between the units in the base register and the units in the source.

The source can after these preparations be included in the Production System and routine editing and adjustment of variable formats can be done. Once it has been included, the source can be used for different purposes. One is to use it as input for a specific statistical product/survey after more register-statistical processing including micro integration with other parts of the system. This new product or register survey should then be judged with the output quality indicators in Chart 2.

The Production System gives opportunities to create new more advanced products by combining existing statistical registers. Alone, the *input data quality* of each component is not sufficient for the new product, but after integration of a number of sources it is possible to create a new product with high output quality. There are many examples of such *integrated registers* at the Nordic NSI:s and they are used mainly by researchers.

2.3 Production Process Quality

Quality of an administrative register or source is defined as the usability of the source for the production of statistics at the NSI.

Apart from using a source directly for producing statistics it may be possible to use a source to improve some part or some parts of the production system at the NSI. This aspect we call the *production process quality of the source*.

At the time of quality assessment of an administrative register, indicators on its present use and its potentials for supporting current statistics production can be measured. As we here primarily discuss enterprise and trade statistics, a list of such parts of the system that could be improved by an administrative source can be given. Can the source be used to …

… improve the Business Register?

… improve the Structural Business Statistics survey (SBS)?

… replace SBS-questionnaires to some extent?

… improve other enterprise surveys?

… improve Intrastat?

… replace Intrastat-questionnaires to some extent?

… improve other trade surveys?

This way of using administrative sources is today often overlooked.

The importance of the NSI's ability to utilize the potentials of a new register can be illustrated by a simple production function example, where registers are treated as a production factor besides capital and labour.

Using a production function with a total cost restriction, the optimal budget shares for the different production factors are functions of the factors unit costs and their output elasticity. This means that, if the production is in an optimal state, the decision on use of a new administrative source will move the production into a non-optimal state. The loss in productivity can be compensated for if the new register can be used in such a way that the output elasticity increases. This means that the NSI must possess the competence on how to utilize the register for increasing the productivity.

The production function example is simple but it illustrates a number of issues to be considered when judging the production process quality aspect of a new administrative data source.

- First, incorporating a new register for producing a set of statistics do not only have effects for the intended use, it also have implications for a large part of the production system at a NSI. Indicators such as the ones specified above helps in determining the benefits of incorporating the register within the production system, benefits that have to be weighed against costs and alternative usage of resources.
- Second, a NSI must possess the competence to utilize the new register, not only for the intended use, but also on the potentials for improving other surveys, both sample and register surveys. Again, the judgment of the indicators exemplified above hinges on the ability to identify potential usage.
- Third, the potentials of using registers in statistics production are restricted by available statistical methodology. Benefits of registers can be increased by developing new and appropriate *register statistical methods*.
- Fourth, it is questionable if the present usage of administrative data at NSIs is at an optimal state. Again, a movement towards a more efficient use of existing administrative data sources is dependent on the ability to develop appropriate survey designs and methodologies.

2.4 Economic Statistics – methodological challenges

In this report we discuss enterprise and trade statistics as the research project[2] we are working with has this demarcation. Even if our principles as a rule are general, it gives us better opportunities to develop good methods if we use the fact that we work with economic statistics. All work with statistical methods should start with a clear understanding of the subject-matter issues that defines the objectives behind the survey. In this research project we use this demarcation in a positive way. We can then relate to the Business Register and the needs of the National Accounts when we study quality issues.

There are some issues that are very important for economic statistics, most variables are quantitative flow variables which makes the accruing problem important. Also the definition of the

[2] The authors work within the European Commission's research project *BLUE-Enterprise and Trade Statistics, Work Package 4: Improve the use of administrative sources*

statistical units is crucial, turnover for enterprise units, legal units or local units are three very different variables. In the enterprise population the units are reorganised continuously and this fact gives rise to important and difficult methodological problems.

3 Comparison of Indicators

How should an administrative register or source be analysed to judge its output quality, input data quality and production process quality? We recommend the following work process with four steps described in sections 3.A-3.D. During each step a number of quality indicators can be analysed. As we delimit our study to enterprise and trade surveys it is relevant to relate to the user needs of the National Accounts when we judge quality.

3.A Information from the Administrative Authority

Tax forms, supporting brochures, handbooks etc. should be studied. This is the first step in the work of analysing an administrative register or source. It is also recommended to interview persons at the administrative authority that is responsible for the source.

Chart 3. Indicators of output and input data quality – relevance

Indicator	Quality factor	Description
A1	Relevance of population	Definition of the administrative object set. Which administrative rules determine which objects are included? Is this set suitable as statistical population?
A2	Relevance of units	Definition of the administrative units. Are these units suitable as statistical units?
A3	Relevant keys	Are there primary keys and foreign keys in the source that are suitable for micro integration within the NSI?
A4	Relevance of variables	Definitions of the administrative variables. Are these variables suitable as statistical variables?
A5	Relevance of reference time	Are reference times suitable for statistical usage? What rules for accruing accounting data between months and years are used?
A6	Study domains	Can the units be allocated between relevant study domains? Are there variables describing domains in the source or can the units be linked with domain variables in the Business Register?
A7	Comprehensiveness	Does the source contain a small/large part of an intended population? Does the source contain few/many statistically interesting variables? Can a small/large number of existing surveys benefit from the administrative source?
A8	Updates	How often and at what time points is the administrative register updated?
A9	Delivery time	Time for deliverance of the administrative register from register holder to the NSI
A10	Punctuality	Difference in time between deliverance and agreed deliverance time point
A11	Comparability over time	Extent of changes in the content of the administrative register over time

3.B Analysis and Data Editing of the Source

The next step in the work with analysing an administrative source is to analyse a set of micro data from the source. Usual statistical description and exploratory data analysis should first be performed – how are different variables distributed etc.? After that, an analysis resembling usual editing should be performed. The aim with this analysis is to diagnose the source and should not be confused with the automatic editing performed during routine production of statistics. A clear understanding of the administrative variables is necessary to create good editing rules. It is recommended that persons from the administrative authority are consulted.

Chart 4. Indicators of output and input data quality – accuracy

Indi-cator	Quality factor	Description
B1	Primary key	Fraction of units with usable identities. The primary key should have correct format and reasonable values.
B2	Foreign keys	Fraction of units with usable foreign keys. Foreign keys should have correct format and reasonable values.
B3	Duplicates in the source	Fraction of identities that occur more than once. Fraction of records with different identities but the records are otherwise identical.
B4	Missing values	Fraction of missing values for the statistically interesting variables.
B5	Wrong values	Fraction of wrong or unreasonable values for the statistically interesting variables.

3.C Integrate the Source with the Base Register

The integration will give three object sets: Units in the source only, units in the base register only and units in both. What does the mismatch indicate? Are there quality problems in the base register or in the source? The units that constitute the mismatch between the source and the base register should be analysed carefully. Here, the base register in question is the Business Register abbreviated as BR in the table below.

Chart 5. Indicators on output and input data quality – accuracy

Indi-cator	Quality factor	Description
C1	Under-coverage in BR	Fraction of units: There are enterprises/units that have been active during the reference period but are missing in the BR or are coded as inactive in the BR.
C2	Under-coverage in the source	Fraction of units: There are enterprises/units that have been active during the reference period according to the BR but are missing in the source.
C3	Over-coverage in BR	Fraction of units: Enterprises/units are coded as active in the BR and belong to a category that is covered by the source, but they have no reported activity in the source.
C4	Over-coverage in the source	Fraction of units: There are units in the source that belong to a category, or seem to belong to a category, that is not statistically relevant.

3.D Integrate with Surveys with Similar Variables

This step is the most difficult one of the fours steps and requires subject matter competence and ability to work with statistical analysis. Special statistical methods are needed to evaluate the indicators below. The source should be matched with all relevant surveys and registers with variables that are similar to those in the source. It is recommended to start simple: Match with the most important survey or register, then add more and more sources into one micro integrated set of data with all sources that should be compared and evaluated.

A source with enterprise data can be matched with e.g. the Structural Business Statistics survey (SBS) that contains at least one similar variable compared with the source. What do the comparisons between variable values from the source and the SBS indicate? Are there quality problems in the SBS and/or in the source? These findings are important indicators of the usability of the source.

An administrative source with export/import data, e.g. the VAT-register, can be matched or micro integrated with the Intrastat survey and also with other trade surveys. Quality problems that these comparisons indicate are important indicators of usability of the source and/or quality problems in Intrastat or the other trade surveys.

In the chart with indicators below, we use the SBS survey as an example of a survey or register that is compared with the administrative source in question. When the source is compared with other surveys or registers, SBS in the chart below is replaced. More indicators than those in the chart below can be developed. However, this requires methodological work and such indicators may be added in the future.

Let us assume that the administrative sources we want to evaluate have been matched with the Business register under step C above and are now matched with the SBS and other relevant surveys/registers. The following quality indicators can be analysed with this micro integrated data set where all relevant sources/surveys can be compared. The indicators can be used to describe quality problems with the Business Register, the SBS survey and/or the administrative sources we want to evaluate. As the SBS as a rule uses the BR as frame, the same errors can be found in both the SBS and the BR.

Chart 6.Indicators on input data and production process quality

Indi-cator	Quality factor	Description
D1	Relevance of variables	Variables in the sources can now be *compared with similar variables* in other surveys/registers. This can reveal that the variable in some source is not as relevant as was assumed with indicator A4.
D2	Relevance of variables	An administrative variable can be used as *auxiliary variable* for a sample survey even if its definition is not adequate enough for direct use in a statistical product. The correlation with sample survey variables should therefore be investigated.
D3	Relevance of variables	An administrative source may contain information that can *improve some parts of the Business Register* (BR). Information on Industrial Activity or Sector even for only a small number of enterprises can reduce missing values in the BR. Information on how administrative units are related can improve the quality of Enterprise Units in the BR.
D4	Under-coverage in the BR and SBS	Fraction of population total by industry: Enterprises/units that have been active during the reference period according to the administrative sources but are missing in the BR and the SBS.
D5	Over-coverage in the BR and SBS	Fraction of population total by industry: Enterprises/units that are coded as active in the BR but have not reported any activity in any administrative source. These enterprises may have been treated as nonresponse in the SBS and non-zero variable values were imputed.
D6	Duplicates in BR, SBS and administrative sources	The same enterprise unit can use more than one legal unit and be represented by more than one record in the BR and the SBS. When reporting to tax authorities different legal units can be used. The SBS may combine questionnaire data from one legal unit and use the same data from a different legal unit. To eliminate double counting aggregate enterprise units can be created. This kind of error can be revealed by analysing the set of data based on all relevant sources/surveys.
D7	Wrong units in the BR and SBS	After profiling work, aggregate enterprise units consisting of many legal units are created in the BR and used by the SBS. Through analysis of the combined set of data based on all relevant sources/surveys it is possible to find that some of these aggregate enterprise units are wrong indicating that more legal units should be added.
D8	Missing values	Undercoverage in the BR will give rise to missing values in all classification variables that should be stored in the BR. NACE, sector, geographical region are important variables that will be missing for units that constitute the undercoverage.

Indicator	**Quality factor**	**Description**
D9	Wrong values or wrong units – cross section data	When data from all sources, surveys and registers are integrated, it is possible to do consistency editing. When similar variables from different sources differ, the reasons behind these deviations should be analysed. The cause can be wrong variable values in some source or different units that have the same identity. It is often the case that different sources are influenced by time in different ways – e.g. the sum of 12 monthly values may not correspond to a yearly value because the unit has changed over time (but the identity is the same).
D10	Wrong values or wrong units – longitudinal data	An important situation arises when data from different time periods are integrated. Here it is important to evaluate the longitudinal quality of both variables and units.

Summarizing the indicators it is seen that the completion of all four steps A-D gives necessary measurements of indicators for assessing output quality, input data quality and production process quality.

4 Examples

In this section we illustrate how the method and indicators in previous sections can be used to analyse the usability of administrative sources.

4.1 The Population Register

A number of administrative sources used by Statistics Sweden are used as a single source for producing statistics. The reason for this is that these sources are considered to be of such a high quality that almost no preparations of the input are needed before use in the statistical production process. One example is the population register produced by the Swedish Tax Agency that has been the source used for Statistics Sweden's Population Register.

When the first statistical version of the Swedish Population Register was created at Statistics Sweden about 40 years ago, this administrative source was considered to have high *output quality* so that it could be used almost as it was for statistical purposes. Relevance, accuracy, timeliness and comparability were judged to be sufficient for direct use in the production of statistics.

Today however, the situation has changed. Many young people today study at universities but are registered where their parents live and also many foreigners come and stay in Sweden without being registered as permanently residing and Swedish young people go abroad without reporting to the Tax Agency.

If the Population Register is integrated with other registers according to the methodology in section 3.D both under coverage (indicator D4) and over coverage (D5) will be clear. Integrating with the Register of University Students will give a picture of the errors in the regional codes for young people (indicator D3). These examples show that the *input data quality* today is not so good in the Population Register.

This example illustrates that a source should be evaluated now and then with the methods proposed in this paper as the statistical usability of the source may change over time.

4.2 The Annual Pay Register

The Annual Pay Register is based on the yearly income verifications from all employers to all employees. Also this source is delivered from the Swedish Tax Agency.

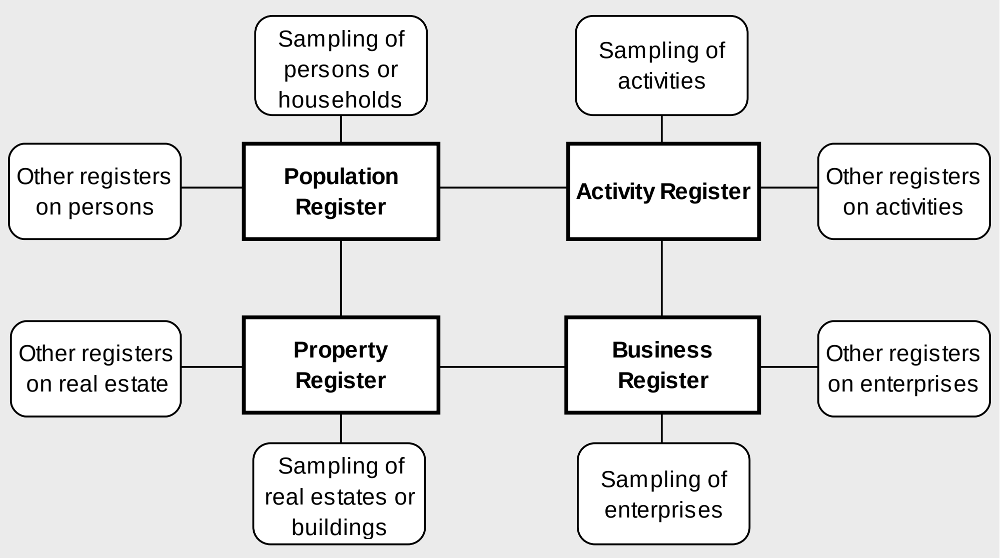
The *output quality* of this source has been checked and it is considered as a source with very high quality. The wage sum definition is very close to what is needed for the National Accounts, accuracy is good and the timeliness is sufficient for the yearly version of the National Accounts.

When the source has been delivered to Statistics Sweden, income verifications for the jobs of employees are transformed into wage sums for enterprise units and local units. Simple editing of these two register versions is performed and then Sector and Economic Activity are imported from the Business Register. Estimates of wage sums by Sector and Economic Activity are produced and delivered to the National Accounts.

4.3 Improving the Production System

The base registers (black rectangles in Chart 7) and the links between them constitute the basis of the Production System in a NSI that utilises administrative data to full extent. All sample surveys use one of them as sampling frame and all register surveys use them as register populations and also use the links in the system to integrate data from different sources. The base registers are discussed in Wallgren and Wallgren (2007, Chapters 2 and 4).

Chart 7. A Production System with full access to Administrative Data


Sampling of persons or households
Sampling of activities
Other registers on persons
Population Register
Activity Register
Other registers on activities
Other registers on real estate
Property Register
Business Register
Other registers on enterprises
Sampling of real estates or buildings
Sampling of enterprises

All administrative sources that can be used to improve a base register – its coverage, its actuality, its classifications as Sector, Economic Activity or Regions, will improve the Production System and the production processes that are used to by different surveys.

Coverage errors are perhaps the kind of "non-sampling errors" that has received less efforts up to now. An efficient use of administrative sources, that results in a coordinated system of statistical registers with base registers of high quality, is perhaps the best way of reducing coverage errors.

According to our experience, almost all administrative sources with enterprise data can be used to improve the Business Register, they have what we here call good *production process quality*. There are also a number of administrative sources with data on persons that today are not used by the Swedish Population Register, but could be used to improve its coverage.

5 Discussion

This paper contains a suggestion of a system of indicators for quality assessment of administrative data. The system consists of a work process in four steps and during each step a number of quality indicators can be analysed.

The system is derived from consideration of three aspects on the usage of an administrative data source – output quality, input quality, production process quality. By the definitions of these three aspects, the indicators of interest are more easily found for output and input quality, while indicators for production process quality are more difficult to develop.

Using a standard system of quality declaration of official statistics, it is possible to define the required information on the administrative data source to assess output quality.

The choice of definition of input quality is less obvious. However, by drawing the line at the stage of receiving a register and a standard check of the contents of the data, the indicators for input quality are more easily identified. The evaluation of the register are here made at face value without relation to use or purpose. Can the register be linked to a base register, are there some technical problems involved when trying to incorporate the register in the register system, to what extent does the register contain missing values and inconsistencies? The *relative* valuation of the register is made either with respect to output quality or production process quality, or both.

What to constitute a set of production process quality indicators is less obvious as this quality aspect involves intended and potential uses of the register. However, for a production process quality assessment of a register, indicators on its present use and its potentials for improving and supporting the statistics production can be measured.

When an administrative source is evaluated with the indicators above, a decision will be taken if the source should be used or not and also how it will be used within the Production System. However, as an administrative register can be used in many different ways, the decision taken will to a great part depend on the ability to relate the source to different potential uses. A greater understanding of the Production System, both its registers and sample surveys, will give a better decision. Also, when the Production System is gradually improved with more and more registers incorporated, then the possibilities to use a specific source will be greater.

The ways different administrative sources are used must be reconsidered now and then as the statistical Production System develops. If administrative sources will be used in an efficient way, then not only indicators as those above are necessary, it is also necessary with a new paradigm – sample and register surveys are not to be looked upon as two different methodologies, they should be treated as two complement alternatives within a single, general survey methodology.

Apart from the statistical methodology aspect, subject matter knowledge and methodology for handling large data sets are important ingredients for a more efficient utilization of administrative data. For instance, for business and trade statistics it is absolutely necessary with a clear understanding of the administrative systems and tax rules that generates the micro data.

Metadata is an important part of the Production System. Not only all statistical products as sample surveys and the statistical registers within the system should be documented. Also all administrative sources that are used, and also sources that have been analysed but not are used at the moment should be documented. This metadata should also be of general and easy availability, promoting consideration of alternative usage of data sources.

One of the first steps in the work with creating a new statistical register, is making an inventory of different sources. All available sources with a connection to the research objectives should be analysed when the new register is created.

This means that it is not sufficient to consider one source at a time, also *combinations of sources* should be considered. By combining sources both relevance and accuracy can be improved. The input data quality of each administrative register or source alone may be insufficient, but a statistical register based on a combination can be of good quality:

- By combining registers with different content, the combined integrated register gives better opportunities of combined analysis of interesting variables and in this way the relevance is improved.
- Turnover in the VAT-register has drawbacks and Turnover in the yearly income statements of enterprises has other drawbacks, but if these two variables are combined it is possible to generate a combined turnover estimate with higher accuracy than each of the two input turnovers.

The system of indicators that we suggest in this paper is based on our present experiences of working with administrative sources at Statistics Sweden. However, the system of indicators and the work process described are also new to us and must be tested with different administrative sources within the Production System at Statistics Sweden.

In the same way that sampling methods are general also register statistical methods should be general. It is therefore also necessary to test the system of indictors and the work process with administrative sources within Production Systems in other countries.

References

Cochran, W.G. (1977). *Sampling Techniques*, Wiley, New York.

Daas, P.J.H., Arends-Tóth, J., Schouten, B., Kuijvenhoven, L. (2008) *Quality Framework for the Evaluation of Administrative Data.* Proceedings of Q2008 European Conference on Quality in Official Statistics, Statistics Italy and Eurostat, Rome, Italy.

Daas, P.J.H., Ossen, S.J.L., Tennekes, M. (2010) *Determination of Administrative Data Quality: Recent results and new developments*. Proceedings of Q2010 European Conference on Quality in Official Statistics, Statistics Finland and Eurostat, Helsinki, Finland.

Daas, P.J.H., Ossen, S.J.L., Vis-Visschers, R.J.W.M., Arends-Toth, J. (2009) *Checklist for the Quality evaluation of Administrative Data Sources.* Discussion paper 09042, Statistics Netherlands.

Dalenius, T. (1969). Designing Descriptive Sample Surveys, In: Johnson and Smith (eds.) *New Developments in Survey Sampling*, pp. 390-415.

Eurostat (2003). *Quality Assessment of Administrative Data for Statistical Purposes; Working group "Assessment of Quality in Statistics"*, Luxembourg, 2-3 October, 2003. Web publication, Eurostat.

Holt, D. (2001). Comment to Platek and Särndal. *Journal of Official Statistics*, **17:1**, 55-61.

Laitila, T. and A. Holmberg (2010), *Comparison of Sample and Register Survey Estimators via MSE Decomposition*, Proceedings of Q2010 European Conference on Quality in Official Statistics, Statistics Finland and Eurostat, Helsinki, Finland.

Särndal, C.-E., Swensson, B. and J. Wretman (1992). *Model Assisted Survey Sampling*, Springer, New York.

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

ISSN 1653-7149 (online)

All officiell statistik finns på: **www.scb.se**
Kundservice: tfn 08-506 948 01

All official statistics can be found at: **www.scb.se**
Customer service, phone +46 8 506 948 01