Quality Assessment in Systems with Registers and Sample Surveys

Anders Wallgren and Britt Wallgren
Statistics Sweden and Örebro University, e-mail: ba.statistik@telia.com

Abstract

The National Statistical Offices (NSO) in the countries in Northern Europe (Norway, Sweden, Finland and Denmark) have developed production systems that are based on statistical registers. Almost all surveys done by the Nordic NSOs are based on these systems of registers. Besides all register surveys that use these registers also sample surveys and censuses use the register system for creating frames, including register variables and as a source for auxiliary variables used for estimation.

These ways of using registers for statistics production are well-known and many countries are developing their production systems to become more and more based on registers as in the Nordic countries. However, the *system of registers* makes it possible to work with quality assessment in a new way:

All registers can be compared at the micro data level and also *all* sample surveys can be compared at the micro data level with *all* registers in the system. Systematic comparisons between samples surveys and registers in the system will give new knowledge of quality in different surveys and also give new possibilities to redesign surveys and improve the quality of the surveys in the system.

Key Words: Administrative data, register survey, survey design, total survey error

1. The transition into a register-based production system

Many countries are today increasing the use of administrative data to produce statistics describing society. Before any administrative registers were used for statistical purposes, the production system was based on maps or address lists and enumerators and interviewers were sent out to interview households and enterprises.

When more and more administrative registers are used the national statistical system is gradually changed into a register-based statistical production system. Sample surveys and traditional censuses are replaced by register surveys that do not require collection of statistical data. Also sample surveys become register-based. The statistical units can be directly sampled from statistical registers – the sample survey design and the estimation methods are improved as register information can be used.

All countries that can do a register-based population and housing census have completed this transition from a traditional production system into a new register-based production system. This transition has consequences regarding survey design and quality assessment, but most people at the national statistical institutes may not yet be aware of these consequences. To understand these consequences it is necessary to fully understand the requirements and possibilities of the *register system* that is the basis of almost all production of statistics after the transition. The register system is described in Wallgren and Wallgren (2007).

The understanding of the role of the register system is today limited – most people at a NSO are fully occupied with their own survey and have little time to study other surveys and make comparisons between related surveys.

2. Survey design in a register-based production system

When a NSO gets access to micro data from administrative registers there can be two approaches to survey design:

With the *traditional approach* we start with the survey content we want. For example, we want to do an income survey and then we start planning for an income register. We search for administrative sources that can be used when an income register is created and develop the methods that should be used.

With the *system approach*, introduced in Laitila, Wallgren and Wallgren (2012), we systematically analyse each administrative source and try to find out *how* it should be

used within the production system or register system. For example, if we analyse income self-assessment from persons we will find that this source can be used in many ways. It can be used for an Income Register and for sample surveys regarding income of households. It can also be used to improve coverage of the Population Register, the Job Register and the Business Register. Also the Structural Business Statistics survey can use this source as there is information regarding sole traders.

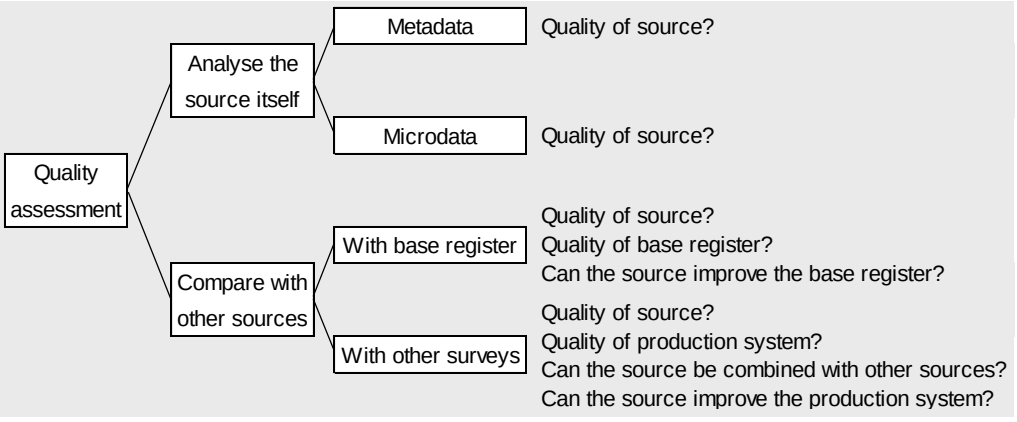
Survey design consists of the efforts we make to maximise the quality of the estimates generated by a specific survey, subject to cost or budget constraints. With quality we as a rule mean accuracy, but also other quality dimensions can be included as relevance, comparability and coherence. Biemer (2010) uses the term *"fitness of use"* for this broader quality concept. The transition from a production system without registers into a register-based system will e.g. reduce the costs for a Population and Housing Census or a Labour Force Survey. It will also be possible to improve quality, census information can be produced every year, and the accuracy of the LFS will be improved when better auxiliary variables can be used.

3. Quality assessment in a register-based production system

Different kinds of survey errors are used as planning criterions when we work with survey design. For the design of sample surveys this planning work is well known and discussed. How should the corresponding planning process for register surveys be structured? In Laitila, Wallgren and Wallgren (2012) we describe the system approach to survey design as consisting of the four steps illustrated in Chart 1. Each administrative source is analysed in the following way:

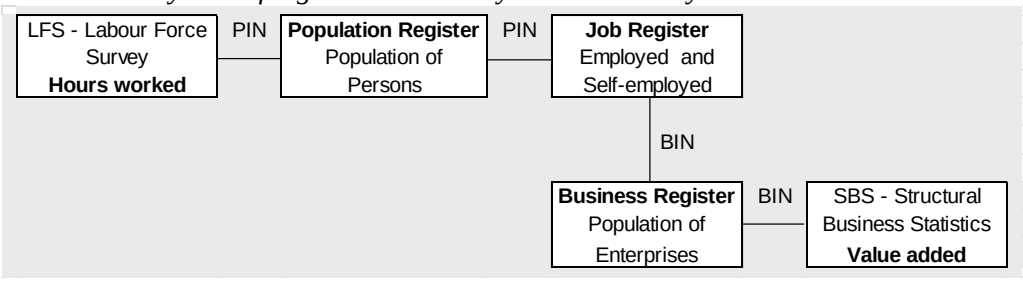
a) Metadata regarding the source is analysed. The relevance is determined.
b) Microdata from the source are analysed. Some aspects of accuracy are determined.
c) The source is compared with its base register (e.g. the Population Register or the Business Register). Some aspects of accuracy of the source and the base register are determined and it is also determined if the source can be used to improve the base register.
d) The source is compared with all surveys in the system containing *similar variables*. Aspects of accuracy of the source and the surveys used for comparisons are determined. It is also determined if the source can be combined with other sources for a new survey and if the source can be used to improve other surveys in the system.

Chart 1. The work with quality assessment of an administrative source



We have tested this system approach to survey design by analysing microdata from five surveys. The intention was to design a new survey where *productivity* by industry in the sector of non-financial enterprises should be estimated with estimates of *value added* from the Structural Business Statistics survey (SBS) and estimates of *hours worked* from the Labour Force Survey (LFS). To analyse the quality of these estimates it is also necessary to analyse the registers that constitute the links between the SBS and the LFS. This means that the Population Register, the Job Register and the Business Register also were analysed. The system is illustrated in Chart 2.

Chart 2. The system of registers and surveys that was analysed



If the object sets in these surveys and registers are compared, undercoverage and overcoverage by sector and industry can be estimated. After comparison of the Population and Job registers we found that the undercoverage in the Population Register due to foreigners working in Sweden is 0.6% of all persons or 1.4% of all employed persons. The estimates of productivity must be corrected for this.

Both the LFS and the Job Register contain PIN, the personal identity numbers. These two datasets can then be matched and Chart 3 illustrates different kinds of errors that were found in the integrated data set.

Chart 3. Example of integrated microdata from the LFS and the Job Register

LFS	LFS	LFS	LFS	LFS	LFS	Job Register		
PIN (1)	Hours worked (2)	Hours usually worked (3)	Sector (4)	ISIC (5)	Weight (6)	PIN (7)	ISIC (8)	Sector (9)
PIN1	12	20	6	56100	32.2	PIN1	56100	110
PIN1	16	20	6	56100	28.8	PIN1	56100	110
PIN1	0	20	6	56100	27.9	PIN1	56100	110
PIN1	20	20	6	56100	33.1	PIN1	56100	110
PIN2	40	40	6	56100	32.4	*	*	*
PIN2	40	40	6	56100	31.5	*	*	*
PIN2	40	40	6	56100	33.2	*	*	*
PIN3	40	40	1	01110	32.1	PIN3	81300	320
PIN4	10	10	6	01110	51.5	PIN4	43320	611
PIN5	45	40	6	01131	40.4	PIN5	01500	611
PIN6	30	30	6	01191	43.1	*	*	*
PIN7	5	8	6	01191	45.7	PIN7	01134	110
PIN8	40	40	6	01199	48.1	PIN8	01430	110
PIN9	60	40	6	64190	47.1	PIN9	55102	212
PIN9	60	40	6	64190	44.7	PIN9	55102	212

The respondents in the LFS are interviewed eight times, once every third month during two years. Each interview concerns the conditions during a specific week just before the interview. We have combined data from the LFS and from the Job Register for 2009. The LFS data describe a sample of the population 15-74 years old and the employment status during one to four specific weeks during 2009 for each respondent. The Job Register describes all jobs for all persons that were employed during the whole year or parts of the year 2009.

In Chart 3 an example with data for a small number of persons is given. The two persons PIN1 and PIN2 both work in restaurants (ISIC 56100) but only PIN1 is found in the Job Register – this is an indication of that PIN2 is a person working in the black sector. PIN3-PIN9 are persons where industry defined by ISIC differs in the LFS and the Job Register. Finally, PIN9 shows also that the Sector variable differs between the two sources. In the Job Register sector 110 and 611 belong to non-financial enterprises, 212 to financial enterprises and 320 to central government. In the LFS the sector code 1 means central government and 6 means non-financial, financial or non-profit sectors – the two sector variables thus differ in definitions, this shows that the sources are not coordinated.

In the example in Chart 3 both the persons PIN1 and PIN2 were interviewed several times during 2009 and each time they were classified as employed in the LFS. However, for only one of these persons preliminary tax has been paid by the person's employer. We suspect that the second person is a person that is occupied in the black

sector of the Swedish economy. An estimate of hours worked by persons of this kind is in the chart below: 0.6% of all hours worked in the LFS 2009.

Chart 4. Hours worked by employed in the LFS, millions per week 2009

ISIC	All hours in LFS	Hours not in Job Register	Not in Job Register %
Agriculture, forestry, fishing	1.129	0.020	1.8
Construction	7.447	0.055	0.7
Wholesale and retail trade	13.536	0.106	0.8
Hotels and restaurants	3.070	0.063	2.1
…	…	…	…
All	115.064	0.706	0.6

At the Swedish National Accounts corrections for black work has been done regarding hours worked. Chart 4 indicates that black work already can be included in the estimates.

Chart 5 illustrates that the Sector coding in the LFS is not coherent with the SBS. This is a typical example of that social statistics and economic statistics often are two separate parts of a statistical office that we have noticed in many countries, not only in Sweden. This difference in sector coding makes productivity estimates based on a combination of the LFS and the SBS impossible if sector in the LFS is not corrected.

Chart 5. Number of employed with one job by sector in the LFS, thousands

Sector according to Job Register:	Sector according to LFS: Private	State	Municipalities	Counties	Unknown	All
Non-financial enterprises	1 848.7	1.7	5.4	1.4	7.7	1 864.9
Financial enterprises	66.6	0.2	0.0	0.0	0.1	67.0
Central government	5.6	148.7	0.3	0.1	0.8	155.4
Municipalities	6.9	0.4	536.3	0.3	1.1	545.2
Counties	1.8	1.0	0.9	160.4	0.3	164.4
Non-profit institutions	59.9	0.2	1.0	0.2	0.2	61.4
Sector unknown	9.8	0.0	0.1	0.0	0.6	10.5
All	1 999.4	152.1	544.0	162.5	10.9	2 868.8

Chart 6 illustrates the problems associated with industry. The target codes are the codes in the Business Register that also are used in the Job Register. In spite of that the Job Register is used when coding industry in the LFS the LFS codes differ from the target. At the two-digit level 11.8% of the employed persons in the LFS that have only one job, have wrong ISIC codes in the LFS (at the 5-digit level 15.9%). This fact makes productivity estimates based on a combination of the LFS and the SBS impossible if the LFS is not corrected. In Chart 6 the industries with the most serious coding problems are included. Some industries have wrong codes for 40% - 50% of the employed persons. The reasons behind these coding problems should be analysed.

Chart 6. Number of employed with one job by ISIC in the LFS, thousands

ISIC	ISIC	ISIC in Job Register	Same code in LFS	Wrong code in LFS Persons	Wrong code in LFS %
Manufacture of beverages	11	3 687	2 146	1 541	41.8
Pharmaceutical products	21	12 227	5 728	6 499	53.2
Computer, electronic, optical products	26	32 366	17 426	14 940	46.2
Wholesale trade	46	142 865	127 928	14 937	10.5
Retail trade	47	175 486	162 891	12 595	7.2
Business support activities	82	32 015	16 766	15 249	47.6
Public administration	84	167 722	149 958	17 764	10.6
Education	85	310 805	286 170	24 635	7.9
...	...	...	...	...	...
	All:	2 868 809	2 530 335	338 474	11.8

Chart 7. Comparing populations in the Business Register (BR) and the Job Register

	Number of enterprises	Gross Pay SEK million
Undercoverage in BR	31 393	6 562
Overcoverage in BR	11 301	
Total population:	331 478	1 241 787

In this chart, the population of active employers is compared with employers in the Business Register.

The undercoverage in Chart 7 above is 9% of the units in the final population, but only 0.5% of the gross pay in the Yearly Gross Pay register regarding all sectors. The undercoverage in the Business Register typically consists of small enterprises.

Chart 8. Undercoverage in the Business Register. Non-financial enterprises

ISIC	Selection of industries	Total gross pay	Undercoverage	
		SEK million	SEK million	%
	Information on industry missing	1 197	1 158	96.7
01	Crop and animal production, hunting	4 843	276	5.7
18	Printing and reproduction of recorded media	5 180	110	2.1
68	Real estate activities	18 350	237	1.3
78	Employment activities	11 867	198	1.7
82	Office support, business support	6 119	112	1.8
95	Repair of computers and personal and household goods	1 196	16	1.3
	All	816 939	5 872	0.7

To be able to correct estimates from the SBS we must have information on under-coverage by economic activity. Chart 8 shows some estimates of undercoverage errors regarding total gross pay. The same kind of estimates of undercoverage errors regarding turnover can be generated if the SBS and the VAT-register are matched.

Chart 9. Undercoverage and overcoverage in the SBS

Legal units that are employers in SBS or the Job Register			Gross Pay, SEK billions 2009	
SBS	In Job Register	Legal units	SBS	Job Register
Not in SBS	Yes	21 392	0.0	5.4
SBS: questionnaire	No	76 137	2.0	0.0
SBS: administrative source	Yes	246 806	543.8	542.0
SBS imputed	No	145 993	3.4	0.0
SBS imputed	Yes	17 805	21.7	19.6
All employers		508 133	570.9	567.0

The total population in the SBS for 2009 consists of 927 904 Kind of Activity Units. 715 of these get a full SBS-questionnaire, for the rest of the population the Yearly Income Declarations are used. In this part of the population the Kind of Activity Units are the same as the legal units used for taxation. The legal units that are employers in the SBS or the Yearly Gross Pay survey are described in Chart 9.

The SBS survey suffers from both overcoverage and undercoverage. The 21 392 legal units with gross pay equal to 5.4 SEK billion that are not in the SBS are undercoverage in the SBS. The 145 993 units that are in the SBS with gross pay equal to 3.4 are overcoverage in the SBS.

These coverage errors in the SBS arise because the population for SBS 2009 was created during November 2009. The YGP 2009 is based on more complete information from September 2010.

The inconsistencies between the two surveys in Chart 9 are small on the aggregate level, but if Chart 9 is disaggregated to show gross pay by industry the inconsistencies for many industries are large.

Conclusions: Quality assessment in a register-based production system

Charts 3-9 show some of the errors we have found when we tested the system approach to quality assessment. The system approach has proved to be important – when we compare many sources and surveys it is possible to detect potential problems in a statistical production system. This is illustrated by the results presented in this report. The traditional way of working is to consider one survey or one administrative source at a time. For both quality and efficiency reasons it is necessary to abandon this tradition and adopt a statistical systems approach as the general method for delivering official statistics.

The errors we have found are serious. We think that similar errors also exist in other countries. However, it is only in a country with a register-based production system that it is possible to find the errors and start the work with correcting them.

4. Total survey error in a register-based production system

The total survey error describes all errors that give rise to lack of accuracy. The sampling error is always measured in sample surveys, but the other non-sampling components are seldom measured. However, the non-sampling errors should always be considered during the design process. The total survey error is discussed by Groves and Lyberg (2010) and is considered to be *"the conceptual foundation of the field of survey methodology".*

Register surveys should also be included in the survey methodology and this area is becoming more and more important as the use of administrative data is increasing. What similarities and differences can be found if we compare the sample survey based ideas in Biemer (2010) and Groves and Lyberg with the example in this paper where all surveys are register-based?

The most important difference is that Biemer, Groves and Lyberg discuss *one* (sample) survey at a time, it is *one* survey that should be designed so that the total survey error is minimised under the budget constraints. In the example above with register-based surveys a *system* of surveys is considered. In our system a sample survey, the LFS, is included, but some survey error components of the Swedish register-based LFS is determined by undercoverage in the Population Register. So we cannot design the LFS alone, we must simultaneously consider the design of the Population Register and other parts of the Swedish production system that is used together with the LFS.

Another difference is that we can measure many important (non-sampling) errors of the LFS and the other surveys in the system. We can do this by integrating data from different parts in the system. We compare the Population Register and the Job Register and find coverage errors; we compare the Job Register and the Business Register and find more coverage errors. And we can compare classification of economic activity in a number of surveys and describe the inconsistencies in the system. We do not have to use *quality indicators*; we can measure relevant quality components directly.

References

Biemer, P. (2010): *Total Survey Error – Design, Implementation and Evaluation.* Public Opinion Quarterly, Vol. 74, No 5, pp. 817-848.

Groves, R., Lyberg, L. (2010): *Total Survey Error – Past, Present and Future.* Public Opinion Quarterly, Vol. 74, No 5, pp. 849-879.

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

Laitila, T., Wallgren, A., Wallgren B. (2012): *Quality Assessment of Administrative Data – Data Source Quality.* Part two of third deliverable of Work Package 4 of the BLUE Enterprise and Trade Statistics project.