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

Frame Populations and Register Populations

Anders Wallgren and Britt Wallgren
Statistics Sweden
anders.wallgren@scb.se; britt.wallgren@scb.se;

Summary
In sample surveys and censuses that are based on the collection of statistical data, the methods used to define and create the frame to a large extent determine the quality of the survey. A base register based on administrative data, such as the Population Register or the Business Register, is as a rule used to create such frames.

The first step in a register-based survey is to define and create the register population. Base registers are used when the register population of one particular register-based survey is created. However, the methods used when frames respective register populations are created differ in many important aspects. As a consequence the coverage errors will be different.

This important difference between sample surveys and register-based surveys is discussed, and examples are given that illustrate the consequences for statistics based on the Business Register to be used by the yearly National Accounts.

Chapter 1. National Accounts and Business Statistics

A large part of the input to the National Accounts consists of different business surveys. These surveys use the Business Register as sampling frame and as a source of important classification variables such as sector and industry. For this reason, the Business Register has been regarded as the basis of economic statistics at Statistics Sweden.

However, these surveys are done at different points in time and are based on different methods. Some surveys are based on for instance questionnaires sent to monthly samples, other surveys are completely based on administrative sources and some surveys, such as the yearly Business Structure Survey, are based on questionnaires sent to a sample and administrative sources are used for the rest of the frame population. Inconsistencies due to differences between business populations will arise due to these different methods and this will give rise to inconsistencies in the data that are delivered to the National Accounts.

The team that is responsible for the National Accounts receive aggregated data, such as production by industry. At the National Accounts they may see that there are inconsistencies in the data delivered, but as they have no direct access to the micro data behind the estimates, they can't find the causes of these inconsistencies. On the other hand, those who deliver aggregated data to the National Accounts usually only work with their own survey and don't usually compare with other business surveys to analyse if there are inconsistencies.

The variable *sector* is used to define different subpopulations and different surveys are responsible for different sectors. The team at Statistics Sweden working with the Business Structure Survey is responsible for the non-financial enterprises and the non-profit household organisations; different statistical agencies outside Statistics Sweden are responsible for financial enterprises, the sector state and the sector of municipalities. This way of sharing responsibility for different sectors may easily lead to inconsistent populations.

Chapter 2. Registers and Time

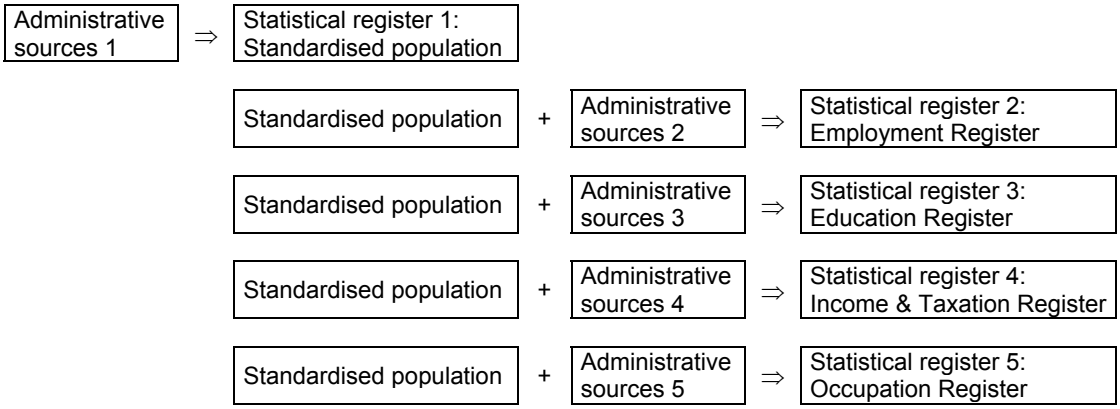
Individual objects, and therefore also object sets, change over time. Objects are born, change location, are altered or cease to exist. These different types of occurrences are called demographic events and it is these that change object sets. When defining a register with regard to time, different register types must be distinguished as discussed in Wallgren (2007). We here mention the following three register types:

1. The *current stock register* is the register version that is updated with all available data on currently active/live objects. The current stock register is used as frame population for sample surveys or censuses. Business sample surveys use currents stock versions created at different points in time *just before* the sample is drawn and questionnaires are sent out.
2. The register referring to a specific point in time, such as the turn of the year, is the version of the register that is updated to describe the object set at that point in time. This update is carried out *after the point in time*, when information on all events up to that point in time is available. This kind of register referring to the population on December 31 is used for almost all register-based surveys on persons at Statistics Sweden.
3. The *calendar year register* is the register version containing all objects that have existed at any point during a specific year. Objects that are added or cease to exist during the year are included with information on the date of the event. It should be used as register population for register-based surveys. This kind of register has not yet been used at Statistics Sweden but is planned to be introduced. The calendar year register for year *t* is created when all administrative data concerning year *t* has been delivered – this may take considerable time, at Statistics Sweden we are able to produce this kind of Business register at the end of year *t+1*.

Chapter 3. Register-based Statistics on Persons

The register-based statistics on persons at Statistics Sweden is created by a decentralised estimation process, but results nevertheless in completely consistent and coherent micro data. This micro data consist of different parts or registers, which are created by different teams at Statistics Sweden. The first step in the process implies that the team responsible for the Population Register creates a *standardised population*. This population is then the basis for those who work with other registers. Chart 1 illustrates the work with register-based statistics on persons.

Chart 1. Decentralized but coordinated processes to create statistics on persons



The standardised population is defined as the population at December 31. The population for December 31, year *t* is created in early February year *t+1*. This standardised population is used as register population in the other statistical registers in chart 1. As the administrative sources 2 – 5 do not overlap regarding statistical variables, the work to create statistical register 2 – 5 can be done independently of each other.

The five registers in chart 1 can be integrated into one register with all variables, which can be used to produce completely consistent register-based statistics on persons. This is illustrated in chart 2 below.

Chart 2. Integrated register based on five registers on persons

PIN	Classification variables from the Population Register					Statistical variables from the …											
						Employment Reg			Education Reg			Income & Taxation			Occupation Reg		
	Age	Sex	Region	var..	var..	var..	var..	var..	var..	var..	var..	var..	var..	var..	var..	var..	var..
1																	
2																	
3																	
•••																	
N																	

This way of basing all yearly register-based statistics on persons on the same standardised register population has been very successful. For many years, Statistics Sweden has produced this perfectly consistent statistics that has been much appreciated by our users.

Chapter 4. Register-based Business Statistics

The present situation of business statistics is a sharp contrast to the situation described above concerning register-based statistics on persons. Instead of receiving completely consistent data, the National Accounts receive data that suffer from inconsistencies. To achieve consistency of business data, it is necessary that the business populations are standardised and coordinated. Inspired by the consistency of our register-based statistics on persons, we are working with a project that aims to investigate how business population consistency can be achieved. We are working with business data for the definite version of the yearly National Accounts. Data concerning year *t* are delivered to the National Accounts during spring of year *t+2*.

First, the population must be defined as the calendar year population that is the population of enterprises active somewhere during year *t*, during the whole year or during a part of the year.

We will illustrate with data from six administrative sources, all regarding the calendar year 2004. The table below shows the number of legal units (LU) in these six sources.

Chart 3. Six sources of a calendar year Business Register 2004

Register regarding 2004		**Number of legal units, LU**
Administrative source 1	Active in Business Register, November frame	872 391
Administrative source 2	Enterprises, yearly income declarations	980 305
Administrative source 3	VAT declarations	689 577
Administrative source 4	Yearly wage sums from statements of earnings	311 943
Administrative source 5	Monthly wage sums	310 628
Administrative source 6	Foreign trade	68 002

The calendar version of the Business Register should be based on *all available* administrative sources 1 – 6. The reason for this is that we want to minimise undercoverage. This calendar year register for year *t* can be used at the beginning of year *t+2*. Even if this register cannot be used as sampling frame, it can be used as register population for all yearly enterprise surveys. Work with the yearly National Accounts regarding year *t* starts during spring of year *t+2*; this means that this kind of register population will also be suitable for the National Accounts.

Chart 4. Calendar year population

	Classification variables from the Business Register		
BIN	NACE	Sector	Region
1			
2			
3			
•••			
N			

Our aim is to create a calendar year version of the Business Register with good coverage and to use this register as standardised population for all business surveys that deliver data to the yearly National Accounts. We also want to use this calendar year register to investigate coverage errors and inconsistencies in the present business statistics.

Chapter 5. Frame and Coverage Errors

Today, the yearly business surveys for year *t* are based on a frame population created during November year *t*. This frame consists of enterprises active during November according to the administrative source 1 mentioned above from the National Tax Board.

Chapter 5.1 Undercoverage

The chart below shows the undercoverage in the Business Register, i.e. the number of legal units missing in spite of the fact that the enterprises have reported to the National Tax Board.

Chart 5. Undercoverage in the November frame of the Business Register 2004

Number of legal units	Adm source 2	Adm source 3	Adm source 4	Adm source 5	Adm source 6
Not in the Business Register	260 943	83 487	33 387	30 533	14 778
In the Business Register	719 362	606 090	278 556	280 095	53 224
Total	**980 305**	**689 577**	**311 943**	**310 628**	**68 002**

The main reason of this undercoverage is that the Business Register is based on administrative source 1, which is early available. This is motivated by the fact that the register is used as a frame for data collection. The November version of the Business Register is used when questionnaires are sent out during January the next year. The administrative source 2 (income declarations) regarding 2004 was available at the end of 2005; source 3 (VAT declarations) was available during spring 2005 for some enterprises but for small enterprises at the end of 2005. Sources 4, 5 and 6 were available during spring 2005. This means that the sources 2 – 6 were not available when the frame was created during November 2004.

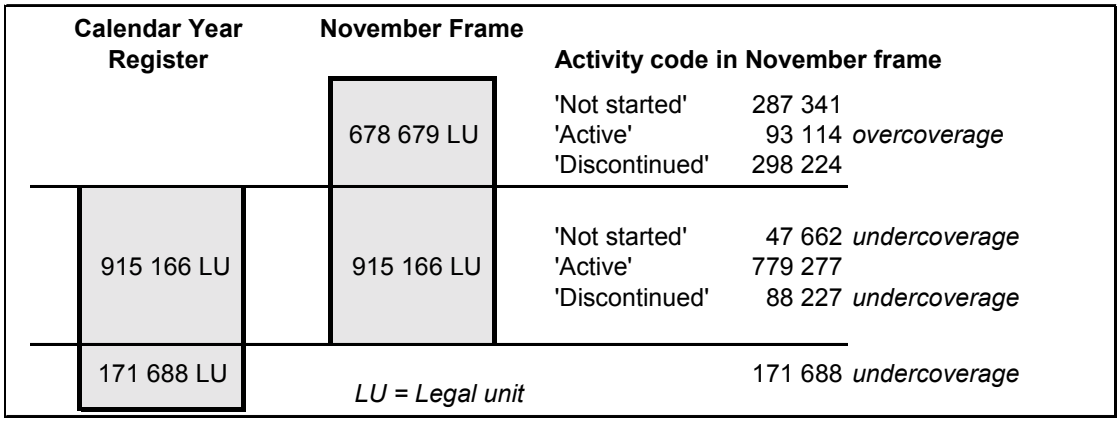
The undercoverage consists of small enterprises and enterprises that existed during a short period of time, e.g. bankrupt's estates and deceased's estates. These enterprises are in many cases small, but as they are many they will contribute substantially to economic statistics, which is shown later in this paper.

Our conclusion is that the calendar year version of the Business Register cannot be based on administrative source 1. This source is the National Tax Board's register with all enterprises (legal units) registered for VAT and/or as employers. Sources 2 – 6 give information about all enterprises regarding actual turnover and wage sums.

Chapter 5.2 The Calendar Year Register compared with the November Frame

We have used the administrative sources 2 – 6 to create Calendar Year Registers. All legal unit identities with economic transactions during each calendar year were included. Declarations where only zeros were reported have been excluded. Chart 6 shows the over- and undercoverage and that the variable 'Activity code' in the Business Register is presently of bad quality.

Chart 6. Over- and undercoverage in the November frame of the Business Register 2004



Chapter 5.3 Coverage errors

Up to now we have only measured over-and undercoverage as number of legal units. For the National Accounts it is the amount of money that counts. In Chart 7 below the overall coverage errors are shown. Most of the enterprises that constitute the undercoverage are small enterprises, in many cases these enterprises are self-employed that combine income as employed with income as self-employed. This is the explanation behind the fact that 53,23% undercoverage in number of legal units corresponds to 2,13% coverage error in the estimate of total turnover. It must also be remembered, that 2,13% is an overall average – coverage errors by industry, are substantial for certain industries (e.g. agriculture).

Also the estimate of *gross annual pay* is 0,65% to low due to coverage errors. This means that even enterprises with employed are found in the administrative sources that are not used by the Business Register today.

The main part of the coverage errors are caused by the category 'Discontinued' in the Business Register. This fact indicates that the rules used by the Business Register today to decide when an enterprise should be classified as 'Discontinued' should be revised.

Chart 7.
Errors due to undercoverage in the November Frame of the Business Register 2004

	Gross annual pay SEK Millions	Turnover SEK Millions	Number of LU
'Active' in the Business Register, gives old estimates	**1 003 186**	**5 582 374**	**779 277**
Missing completely in the Business Register	639	5 061	171 688
'Not started' in the Business Register	177	1 734	47 662
'Discontinued' in the Business Register	5 671	112 363	88 227
Total in Calendar Year Register, new estimates	**1 009 673**	**5 701 532**	**1 086 854**
Coverage error due to undercoverage, new - old estimate	6 487	119 158	307 577
Coverage error due to undercoverage, per cent	0,65	2,13	39,47

The *overcoverage* of 93 114 legal units that is shown in Chart 6 will also give rise to coverage errors. This category is today treated as non-response in the Business Structure Survey. The errors due to these, results in that gross annual pay in the Business Structure Survey is overestimated with SEK 25 500 millions and production is overestimated with SEK 150 180 millions. There is no reason that errors caused by undercoverage and overcoverage will cancel out – as the distribution by industry will differ.

Chapter 5.4 Sector

Sector is an important variable in the National Accounts. In the Swedish National Accounts we distinguish between the following sectors:

– non-financial enterprises,
– financial enterprises,
– Government authorities,
– municipal authorities,
– social insurance, and
– private non-profit organisations.

Many surveys define their populations by sector and totals of many variables are reported by sector. *All* economic active enterprises (and organisations) should contribute to estimates in the National Accounts. Each enterprise should contribute *once*, and no active enterprises or organisations should be *overlooked.*

The enterprises and organisations belonging to the calendar year populations that we have created have a sector code, either taken from the Business Register (present, earlier or later versions) or coded by us with available administrative data.

With the calendar year register, where all units have a sector code, we have checked populations by sector of a number of business surveys. We are looking for inconsistencies regarding populations:

– Are some enterprises included in more than one survey?
– Are some enterprises overlooked by all surveys?

Our analysis shows that the populations of the surveys are not coordinated and consistent. Some enterprises are included in more than one survey and some enterprises and organisations are overlooked by all surveys. In the chart below coverage errors in the survey of Government authorities are shown as an example of inconsistencies regarding populations.

Chart 8. Coverage errors regarding the population of Government authorities

	Gross annual pay 2005 according to…		
	1. The survey of Goverment SEK millions	2. Calendar year register SEK millions	Number of legal units
LU in both (1) and (2)	69 607	70 008	221
LU only in (1)	245	-	8
LU only in (2)	-	1 739	108
Total	**69 852**	**71 747**	**337**

Chapter 6. Frames and Calendar Year Registers

Our conclusions from our work with the Calendar Year Register are the following:

1. If you want consistent estimates from different surveys, all surveys must be based on the same standardised population.
2. If you want good coverage, this standardised population must be based an all relevant administrative sources.
3. This standardised population can be created first after that all relevant administrative sources are available.
4. The Calendar Year Business Register based on all relevant sources should be used as this standardised population for all business surveys used by the yearly National Accounts.
5. The corresponding frame population based on the current stock register can't be used as standardised population as the coverage errors are not under control.
6. The administrative sources will be consistent with the Calendar Year Register. Sample surveys originally based on the frame population can be calibrated against the Calendar Year Register to obtain consistent revised estimates.
7. The methods used to create frames respective Calendar Year Registers are different.
   A statistical office should not have only one Business Register. One version should be the Current Stock Register that should be used for sample surveys giving preliminary estimates.

The Calendar Year Register is an other kind of Business Register that should be used for all register-based business surveys and for revised and consistent sample survey estimates.

8. Never use only one administrative source! To be able to judge the quality of an administrative source it should be combined and compared with other sources. It is to be noted that no one of the sources 1 – 6 mentioned in Chart 3 above is of good quality alone. It is only when they are integrated that we can create a register with good coverage/quality.

References:

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