"Science Stays True Here"
Journal of Mathematics and Statistical Science, Volume 2016, 208-216 | Science Signpost Publishing

Frames and Populations in a Register-based National Statistical System

Anders Wallgren and Britt Wallgren
BA Statistiksystem AB, formerly of Statistics Sweden

Abstract

In this paper we discuss how coverage errors in statistical registers can be measured and how estimates can be corrected for coverage errors by a dual frame approach and calibration of weights.

Keywords: Coverage errors, dual frames, calibration, administrative registers, register surveys

Introduction

Many countries have started to use administrative registers for statistical purposes. When the national statistical institute starts to use administrative registers, the survey methods will change: the new registers will be used for both censuses and sample surveys as frames and as sources of statistical variables. Gradually, the statistical system will become more and more register-based. The countries in North Europe have made this transition from a traditional area frame based statistical system into a completely register-based system.

In such a system, all *sample surveys* are based on frames that have been created by statistical registers. Also, for a *census* as the Farm Census, the Farm Register is used as frame and questionnaires are sent to all holdings to get additional information that is not already in the register. A third kind of survey is based entirely on microdata in the system of statistical registers. The traditional Population and Housing Census is replaced by a number of such *register surveys* that are based on statistical registers created by the national statistical institute. Instead of *frame populations* the register surveys are based on *register populations* created in a different way than frame populations, see Wallgren and Wallgren (2014, p. 131). In a register-based statistical system there are new ways of improving consistency and coherence.

Corresponding authors: Anders and Britt Wallgren, BA Statistiksystem AB, Norrhagev. 30, 71993 Vintrosa, Sweden. E-mail: ba.statistik@telia.com

Coverage errors can be analysed and reduced by combining data from different sources.

However, in many countries the coverage of the administrative systems are not perfect. In developing countries there can exist informal sectors especially in rural regions where persons and holdings are not included in administrative registers. It will then be necessary to combine register-based surveys with area-based sampling. Also, some administrative systems are not updated regularly. As a result, important variables in the registers, as e.g. residential address, can be wrong. The magnitude of such errors can be measured with area frame based sample surveys.

We think that coverage errors is a problem that should get more attention. There is a long tradition to work with sampling and nonresponse errors and the rising nonresponse rates in Statistics Sweden's LFS have recently been discussed in Swedish media. But coverage problems have never been discussed in media. The Population Register used by Statistics Sweden has an overcoverage and undercoverage that both are between 0.5%–1%.

The Business Register used for yearly economic statistics had for the year 2004 overcoverage of about 8% and undercoverage of about 28% regarding number of units and about 2% regarding total turnover (Wallgren and Wallgren (2014) p. 139–140, 222). These coverage errors are also present in all sample surveys that use these registers as sampling frames.

Even if the coverage problems are considered small at the total level, the errors can be highly selective. An example of this is that the overcoverage in the Swedish Population Register generates serious problems for statistics by country of birth. The problem has not yet been tackled.

In a country like Sweden, with many good administrative registers available, these coverage errors can be reduced by using more sources and by improving the statistical methods. In developing countries, however, that have recently started to use administrative registers we suggest that coverage errors are estimated by using dual frames, i.e. combining registers with sample surveys based on area frames. This dual frame approach for sample surveys is discussed by Ferraz (2015) and Carfagna and Carfagna (2010). We will also use the calibration technique developed by Deville and Särndal (1992) to correct estimates for coverage errors.

Coverage errors in censuses are often estimated by post enumeration surveys. The scope and methods used for these surveys differ from the methods we discuss here. In a post enumeration survey, repeated measurements are made using the same frame as in the census. The measurement errors that are found are used to estimate undercoverage errors in the census. A recent paper by da Silva et al. (2015) describes how this method has been used for the Brazilian Census 2010.

Coverage problems in administrative sources are discussed in a recent issue of Journal of Official Statistics. Bakker et al. (2015) give an introduction and describe how capture-recapture methods are used to estimate undercoverage. Calibration methods can also be use to correct for undercoverage. However, the methods assume that there is no overcoverage.

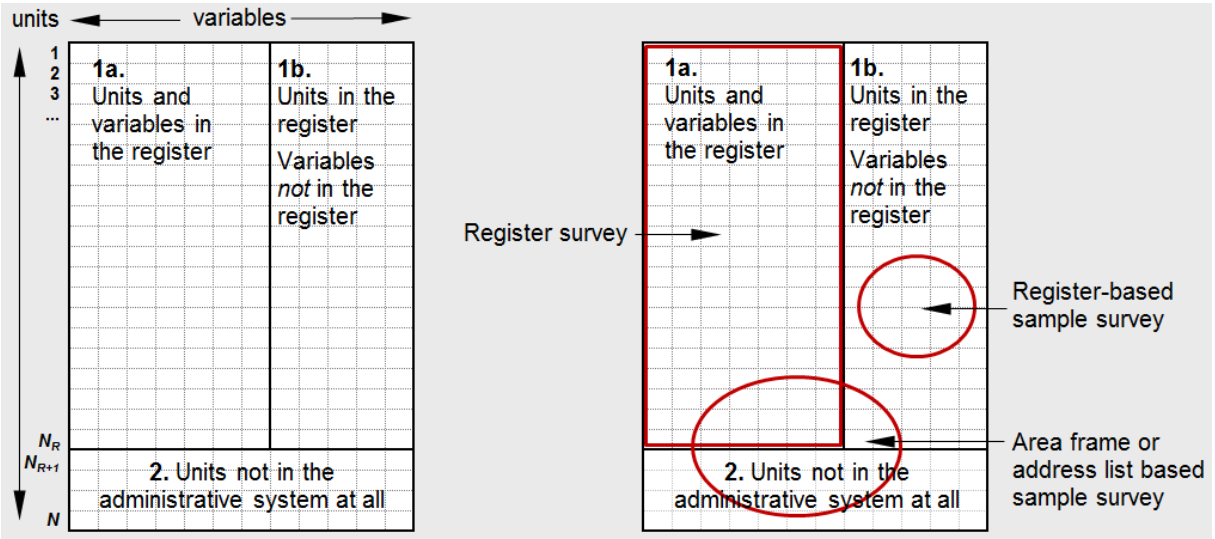
Defining the Statistical Problem

When a statistical register has coverage errors due to undercoverage it can be combined with a sample survey based on an area frame. As the area frame based sample theoretically has no coverage errors the estimates can be unbiased. In a similar manner, a register-based sample survey that uses a register with undercoverage as sampling frame can be combined with an area frame based sample.

In Chart 1 the frame population consists of N units and can be divided in two parts: Part 1 (1a and 1b) consists of the units that are included in the register and part 2 consists of the units belonging to the register's undercoverage. The register survey in Chart 2 can be combined with an area sample and the register-based sample survey can also be combined with an area sample.

Chart 1.
Three different parts of the population

Chart 2.
Three different surveys



From Wallgren and Wallgren (2014), p 227

Ferraz (2015) considers the case when a list-based (or register-based) sample and an area frame based sample containing the same variables are combined during the estimation stage. All variables are collected for both samples. In this paper we prefer to use the term *register* instead of *list*.

We consider here a different case that we think is important for countries that have started using

administrative registers to create statistical registers. Coverage errors and problems with residential addresses that have not been updated may make it necessary to combine a statistical register with an area-based sample. The statistics produced is mainly based on the register or on sample surveys based on frames created with the register. With the register it is desirable to produce estimates by municipality. The area-based sample is used to estimate coverage errors on the national and regional levels and for different categories of municipalities and persons or households. It can also be possible to adjust or correct for coverage errors with calibration conditions based on some register variables.

Assume that we want to produce regional statistics with the Population Register created by a national statistical institute. In Chart 3 some of the variables in the register are listed. The municipalities have been classified in different categories based on degree of urbanization and perhaps also the relative size of an indigenous population. We suppose that this variable can influence the size of the coverage errors for a municipality. It may also be possible to classify households or persons in different categories with different propensities of moving to another residential address. Age, level of education and economic activity are variables that can be used here.

Chart 3. Variables in the three different surveys

Variables in the register	Variables in the area frame based sample	Variables in the register-based sample
Identity number, *Idnr*	If the unit is not in the register: *Idnr* = missing	Identity number, *Idnr*
Residential address	Other variables: same as in the register	Other variables: same as in the register
From date of residence	Possible to add statistical variables	and statistical variables for the survey
District		
Municipality		
Category of municipality		
Gender		
Age		
Level of education		
Economic activity		
Category of person or household		

For an area sample, the same variables as in the register are collected by interviewers. We will find persons in the sample that are not in the register or have different address in the register. In this way we will find undercoverage and overcoverage on different regional levels: wrong municipality, wrong district and not included at all in the register.

Estimation of Coverage Errors

A number of important quality measures can be calculated based on comparisons between the area sample and the register: percentages of persons with wrong residential address, the delay in reporting a new address and undercoverage rates on the regional and national levels. The long-term work with

improvements of the national registration system should be monitored with these measures.

Chart 4 illustrates different kinds of coverage errors. On the national level the area sample gives an estimated undercoverage of 72 000 persons in the register. The area sample estimate is that 80 000 persons are actually living in district A, but they are registered as living in district B. These 80 000 persons are an estimated *undercoverage* of the subpopulation living in district A and at the same time an estimated *overcoverage* of the subpopulation living in district B. Persons that have left the country is another category of overcoverage that cannot be detected with the methods discussed here.

Chart 4. Examples of coverage errors, area sample estimates of the population by district

		District according to register data					
	District	A	B	C	Not in register	All	*Undercoverage*
District	A	2 870 000	80 000	110 000	19 000	**3 079 000**	209 000
according to	B	16 000	3 110 000	40 000	27 000	**3 193 000**	83 000
area sample	C	2 000	10 000	1 530 000	26 000	**1 568 000**	38 000
	All	2 888 000	3 200 000	1 680 000	72 000	**7 840 000**	330 000
	Overcoverage	18 000	90 000	150 000			

Dual Frame Estimation

We assume that the area sample can give us reliable estimates at the regional level and that municipalities define the desired regional level for the estimates based on the register. Estimation with the dual frame approach has some shortcomings that will make the dual frame estimators discussed in Ferraz (2015) difficult to use for the case we consider here:

- We don't think that the area sample can be designed so that dual frame estimates for municipalities can be produced. Cost limits will make this impossible.
- There will be inconsistencies between register data and measurements made by the interviewers. For example some persons will say that they have completed secondary school but according to school register data they have not.
- Regional register data and register-based sample surveys will be influenced by overcoverage errors. The dual frame estimators will only correct for undercoverage. Overcoverage in the register will be detected in the area sample when the interviewers meet persons that have moved *to* addresses in the sample. Regarding persons that have moved *from* addresses in the area sample the interviewers will understand that these addresses are wrong as residential address of these persons, but it will be difficult to determine if the district and/or municipality is wrong or not.

Our conclusion is that the dual frame methods discussed by Ferraz (2015) are not suitable when we have both overcoverage and undercoverage for important domains of study. Errors in the register (list frame) regarding the classification variables that are used to define the domains of study will generate these kinds of coverage errors. Location of residence of persons and economic activity of establishments are examples of such classification variables. If we want to produce statistics based on the Population Register or the Business Register and correct for coverage errors we need a different kind of dual frame approach.

Adjusting for Coverage Errors through Calibration of Weights

The alternative dual frame approach we discuss in this section uses methods that are similar to the calibration method we use to correct for nonresponse in sample surveys. The area sample discussed in the previous sections can be used in two ways:

- To give estimates of errors regarding coverage and errors in basic classification variables in the register as described in the previous section.
- To give estimates that can be used as calibration conditions that can be used to adjust or correct for coverage errors.

Chart 5 shows the structure of the Population Register. The variables in columns (2) – (6) can be used to define calibration conditions if we want to adjust for coverage errors. If we choose to use only *District* and information on the number of persons from each district as calibration conditions, the original weights d_i in column (7) will be transformed into the calibrated weights w_i in column (11).

Chart 5. Seven records from the population register

Identity number (1)	Sex (2)	Age (3)	Municipality (4)	Category (5)	District (6)	d_i (7)	x_1 (8)	x_2 (9)	x_3 (10)	w_i (11)
100002	2	99	1130	2	B	1	0	1	0	0.998
100003	2	33	1224	4	C	1	0	0	1	0.933
100012	1	73	1240	2	C	1	0	0	1	0.933
100013	1	19	1134	3	B	1	0	1	0	0.998
100019	1	64	1243	4	C	1	0	0	1	0.933
100023	1	24	1040	2	A	1	1	0	0	1.066
100026	1	72	1025	1	A	1	1	0	0	1.066
...	...	...	...	...	...	...	...	...	...	...

If we sum columns (7) and (11) by district we will get the original uncorrected estimates and the corrected estimates:

Chart 6. Uncorrected and corrected estimates

District	Original estimates	Corrected estimates
A	2 888 000	3 079 000
B	3 200 000	3 193 000
C	1 680 000	1 568 000
All	7 768 000	7 840 000

In this case it is easy to calculate the calibrated weights. The weights w_i for the records with District A in the register will be 3079/2888 = 1.066 and the weights for the other districts in a similar way. With columns (8) – (10) the same weights can be derived with matrix algebra.

The data from the area sample should be analysed and differences regarding coverage errors for different categories of sex, age, category of municipality and district can be used to build a better model that can be used to adjust or correct the estimates for coverage errors. This methodological work will be very similar to the work we do to adjust for nonresponse in sample surveys.

If there are more statistical registers with data on for example education, employment or income, these registers should use the population defined by the Population Register as their register populations. The calibrated weights w_i should be used for estimation also in these registers. In this way the registers will be consistent and the statistics produced with all these registers will be coherent.

Sample surveys that use the Population Register as sampling frame should also be corrected for the coverage errors in the Population Register. When adjusting the weights used for producing the estimates for the sample survey the estimates are corrected for both nonresponse and coverage errors if register totals that have been corrected for coverage errors are used. In this way we combine the information in the area sample with the information in the register-based sample to obtain estimates that have been corrected or adjusted for coverage errors. With the terms used in Ferraz (2015) we have linked an area frame with a list frame to produce the best possible estimates. However the statistical problem differs from the problem studied by Ferraz, as we consider register-based statistics and use the area sample to correct the register estimates.

Discussion

Many countries have Population Registers, Business Registers and Farm Registers that have both overcoverage and undercoverage problems. Basic register information as residential address or economic activity can be old and incorrect. These quality factors should be monitored by area sampling. It is also possible to improve the quality of the register-based estimates by calibrating weights. Variables in the registers should be used to define calibration conditions that will reduce the coverage errors. In the long run, the national registration systems should also be improved so that the coverage problems gradually are reduced.

Today, we publish estimates from sample surveys with nonresponse rates that are 40% or sometimes more. We can do this as we trust our methods for adjusting estimates for nonresponse with the information in the registers we use as frames. Could we also adjust register-based estimates when we have 40% undercoverage? For us this idea is new and chocking. As statisticians from Sweden we are used to a situation where undercoverage is at most 0.5% – 2%, and that this undercoverage is so small that it can be neglected. First we want to make clear, that we now understand that undercoverage errors can be very selective and that overcoverage or undercoverage between 0.5% – 2% should not be neglected as serious errors can be generated. We also remember that when we studied statistics as young students we were taught that if the nonresponse rate was more than 5%, the sample survey was completely worthless! Our conclusion of this is that attitudes towards errors in surveys change over time. If we have trust in our methods for adjusting estimates for nonresponse and/or undercoverage we can do corrections and publish. Correcting estimates for nonresponse and correcting for undercoverage with the method we propose in this paper requires very similar methods. Our recommendation is therefore:

- combine register data with data from an area frame based sample,
- analyze the coverage problems and define suitable calibration conditions,
- correct estimates and publish if you are confident with your correction method.

References

Bakker B.,Heijden, P. and Scholtus S. (2015) Preface to volume 31, Issue 3 (Sep 2015).

Carfagna E. and Carfagna A. (2010) Alternative sampling frames and administrative data; which is the best data source for agricultural statistics? In Benedetti, Bee, Espa, Piersimoni (Editors), Agricultural Survey Methods, Wiley, Chichester, UK.

Deville, J. and Särndal, C-E. (1992) Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376–382.

Ferraz, C. (2015) Linking Area and List Frames in Agricultural Surveys. Paper written for FAO.

Silva A., Soares M., Carneiro D. (2015) Assessing coverage of the 2010 Brazilian Census. Statistical Journal of the IAOS, 31, 215-225.

Wallgren, A. and Wallgren, B. (2014) Register-based Statistics. Statistical Methods for Administrative Data. Second edition, Wiley, Chichester, UK.