Doing a Register-based Census for the First Time: The Swedish Experiences

Martin Axelson[1], Anders Holmberg[1], Ingegerd Jansson[2]
Peter Werner[3] and Sara Westling[3]
[1]Statistics Sweden, Dept. of Research & Development, SE701 89 Örebro, Sweden
[2] Statistics Sweden, Dept. of Research & Development, Box 24300, SE104 51 Stockholm, Sweden
[3] Statistics Sweden, Process Department, Methodology unit, SE701 89 Örebro, Sweden

Abstract
After years of preparation by several involved agencies and organizations, Sweden's first fully register-based census was conducted in 2011. Statistics Sweden is responsible for the quality of the census statistics, and this paper addresses the methodology and preliminary results from studies which were designed to assess the quality. A system of registers is being built to make continuous production of official register-based statistics on demography, employment, occupation, living conditions and housing possible. Most of it concerns the population of individuals, but register-based statistics on household and family populations as well as the population of buildings and real properties will also be feasible. The quality of the data sources used for new statistics on households is of particular interest.

Key Words: Register-based Statistics, Census 2011, Quality evaluation

1. Introduction

Last year, Census 2011 was conducted in all 27 EU member states. The methods of data collection varied greatly between countries, with Sweden being one of few countries conducting a completely register based census. The system of registers built to facilitate the production of census statistics will in addition be an important part of the future production of official statistics on households and housing in Sweden, in particular allowing for new statistics on households and household composition. Assessing the quality of the household data is thus of great importance.

Although experienced in register-based statistics, Statistics Sweden faces new challenges with a completely register-based census. Register-based statistics differ from traditional surveys in many respects. A register-based survey utilizes a register created and maintained for administrative purposes. In a register-based system, a 'statistical register' is first created where the administrative data are edited and transformed to best meet the aims of multiple surveys. Other typical processes include merger of several administrative registers where issues such as statistical matching and derivation of new variables have to be addressed.

In order to assess some quality aspects of the household data, a sample survey was conducted during the first part of 2012. Before giving a description of the design of the evaluation study and some preliminary results, the sampling design and some quality

issues of a register based census is briefly outlined. The paper concludes with highlights of important issues and a few lessons learnt so far.

2. A Register Based Census

The statistical registers that are used in the census are mainly utilizing administrative data from the Swedish Tax Agency and the Swedish mapping, cadastral and land surveying authority. The system of statistical registers relies on three base registers kept by Statistics Sweden: the Business Register, the Total Population Register (TPR), and the Real Property Register (RPR). The base registers are linked to various subject matter registers such as registers of employment, occupation, education, and buildings. These registers have existed for a considerable time, but still a complete system of registers has not been possible until now.

The system of statistical registers is dependent on the possibility to uniquely link information to objects. Comprehensive identification of persons and businesses is an essential part of the system. Every Swedish citizen has a personal identification number, widely used in all major administrative registers. The personal identification number links an individual to data on for example occupation and education. The personal identification number and the identification number of businesses give the prerequisites for creating statistics on employment. Businesses and individuals are linked to the RPR by the address of the house or building where they reside. For those living in houses the address is unique. For multi-dwelling buildings, the address until recently would only tell the entrance and possibly the floor, but not the flat. Register information also exists on marital status and parent – child relations, but since cohabiting without being married is common in Sweden, this information is not sufficient for forming households. Thus it has not been possible to form households based on administrative data only. With a new register of dwellings, a unique identification number has been given to all dwellings (flats and houses), and there are now new opportunities to create statistics on households and housing conditions.

Constructing a register of dwellings is a cumbersome procedure. The process which has been used in Sweden differs between two types of dwellings, those in one/two dwelling and semi-detached houses, and those in multi-dwelling buildings. For the first type of dwellings, updated addresses in the RPR are used to construct a dwelling id-key, automatically linked to all people registered at the same dwelling address in the TPR. For multi-dwelling buildings the register construction process is very briefly outlined in the following steps:

1. The local administration checks and updates all addresses within a municipality and issues formal addresses for example to dwellings that have only traditional, informal names or to houses sharing one address.
2. The property owner labels flats in a block of flats and submits the dwelling numbers to the land surveying authority where the administrative register of addresses and dwellings is kept. The labeling should be done according to specific rules since the numbers carry information about the ordering of flats on a floor.
3. The property owner informs the residents of their dwelling numbers.
4. By a mailed questionnaire to every adult in the country, the Tax Agency asks individuals about their address of residence, including dwelling number.

5. Individuals living in apartments are expected to be aware of their dwelling number and to give this information to the Tax Agency where the administrative population register is updated.

The population registration on addresses and dwelling numbers started in September 2010 and the last forms where sent out by the Tax Agency in March 2011. All steps are subject to error. The dwelling number is expected to be particularly prone to be missing or incorrect. For further reference, see Axelson et al (2010) or Hedlin et al (2011a, b).

According to the Tax Agency, the data collection resulted in 96.8 percent of the Swedish population being registered on a dwelling number. The goal was to reach 95 percent in each of the 290 municipalities of Sweden. Only eight municipalities did not reach the goal. The Tax Agency continues to improve the register and the goal for 2012 is to reach at least 97 percent in each municipality. In the statistical registers used at Statistics Sweden for the Census, the result is 95.6 percent (i.e. the resulting proportion of individuals registered on a dwelling number when the TPR is merged with the RPR). After adjusting for suspected over coverage, there are 384 000 persons (about four percent) that do not have a registered dwelling number. This means that no dwelling id-key exists and thus there are no linking possibilities between the Dwelling Register and the TPR. The missing data are unevenly spread across the country, with large municipalities being particularly problematic.

3. Measuring Quality

Assessing the quality of a register based survey poses problems that differ from those encountered when measuring quality of sample survey data. Typically, variance based indicators of quality cannot be used, and thus other types of indicators have to be defined and measured.

A main difference between the two types of surveys is that with a register-based survey, data collection is primarily devised for administrative purposes which may differ from the statistical purposes. Hence, effects that the statistical agency has to be aware of might include inadequate definitions of variables and reference periods, or lack of relevance and validity. Data transformed from administrative to statistical purposes should meet the aims of multiple surveys and be fit for many different purposes, such as being the primary source of data collection and ad hoc aggregates, serve as a sample frame, or be used as auxiliary information in estimation.

As of yet, there is no unifying theory for administrative data as a source of data collection, but there is a clear notion of the necessity of such theory. Some references are Wallgren and Wallgren (2007), Eltinge (2010), and Zhang (2011, 2012).

For the Census 2011, Eurostat requires that the quality of the final statistical output is documented. For the Swedish census, the quality of the Dwelling Register is of utmost importance since this is the only part of the system of registers that has not been in use previously. In order to assess the quality, two studies are to be carried out comparing the register data with household data and housing data, respectively. In the next section, the sample survey for evaluation of household data is described. An evaluation of the housing data is planned for the autumn of 2012.

4. Evaluation study

To evaluate the quality of the Dwelling Register and its effect on household statistics, Statistics Sweden has carried out a sample survey of individuals from the TPR. The idea is to get additional observations by asking individuals and compare to what extent their answers are consistent with the register information at the reference time of the census. Following the data collection of the survey, we will try to resolve the true circumstances by analyzing the information from register sources and the survey. Based on the analysis of data and subsequent processing (reconciliation), we will estimate some quality measures such as the net and gross errors of categorical register variables.

The evaluation survey does not only aim for quality measures of the household data in the registers used for the census. To a certain extent, the design of the evaluation survey is also tailored to pick up information which can be used to improve census statistics. Groups of individuals with characteristics where we know data are partially missing in registers, or where we suspect a lower quality in the register, are targeted in the sampling design. If the evaluation survey is successful, data about these groups can be used as auxiliary information in census calculations.

Due to cost restrictions, the length of the survey questionnaire and the data collection had to be limited. Of the census variables only some carefully chosen vital register variables are included in the survey and as far as data collection is concerned, a sequential mixed-mode data collection with the order web-paper-CATI has been used.

4.1 Description of the Sampling Design

A stratified simple random sample of 15 000 individuals was drawn from the TPR. A total of 108 strata were defined by five variables, with categories as defined in Table 1.

Table 1: Variables used for stratification

Variable	*Categories*
Dwelling id-key exists in TPR	Yes No
Municipality of residence (combined with dwelling id-key exists in TPR)	Stockholm region (Yes) Göteborg region (Yes or No) Malmö region (Yes or No) Municipality with more than 70000 inhabitants excl. Stockholm, Göteborg, and Malmö (Yes or No) Other municipalities (Yes or No) Stockholm county area (No) Norrtälje (No)
Age class	18-34 years 35-74 years 75 years or older
Type of dwelling	(semi)-detached house Apartment block Communal establishment, not dwelling house, or missing value

Number of families in dwelling according to TPR	1-2 3 or more No dwelling id-key

The variable Municipality (combined with dwelling id-key exists in TPR) is essentially the size of the municipalities (the three largest cities in Sweden, other fairly large municipalities, and small municipalities). The combination with the existence of dwelling number is needed because the Stockholm area has a high proportion of missing dwelling identification numbers, and in particular the municipality of Norrtälje is problematic in this respect.

The number of families in dwelling according to TPR defines categories where there is likely to be false values in the register.

The sample was allocated to strata by a two-step procedure. The first step was to allocate approximately 1/3 of the sample to the group with no dwelling number in the TPR. The motive to oversample the individuals with missing dwelling numbers is to collect more certain information about the characteristics of this group. The group is of particular interest since we suspect its members to have missing values, both in register based statistics and if they are selected in sample surveys. In the second step a sequential scheme with proportional allocation using both minimum and maximum constraints in each stratum was used in each of the two groups in step one.

Data collection started in January and ended in May 2012. Respondents were given the opportunity to answer online or by a mailed questionnaire. Interviews as well as follow-up for reconciliation were conducted by telephone.

The respondents were asked to confirm the address they were registered at on 31 December 2011 (reference date of the census). If the address given was incorrect, they were asked for a correct address. Other questions concerned if the dwelling is owned or rented, how many dwellings there are at the same address, and how many other persons were living at the same address on the reference date. For all others living at the same dwelling, name, sex, year of birth, and whether the person in question was living together with parent or spouse/partner was asked for.

4.2 Issues

Several issues have to be considered. An important question (almost of a philosophical nature) to consider when quality is to be assessed by means of an evaluation study is what to view as correct information, if anything. Another issue that has to be handled is missing data in the evaluation study.

4.2.1 Reconciliation

The evaluation of the register has itself to be evaluated. Which data are correct, the sample data or the register data, or no data? Discrepancies can differ in magnitude, how large a discrepancy can be accepted before it should be viewed as a serious flaw to data? When discrepancies are serious enough, there must be methods to investigate and, if possible, correct them.

The following are examples of questions of basic interest for the evaluation:

- The proportion of incorrect or partly incorrect addresses. In particular, the proportion of false dwelling identification numbers is of interest.
- The proportions of persons or households renting or owning their dwelling
- Estimated deviations from the register of number of households in a dwelling
- Estimated deviations in household composition between register and evaluation survey

Household composition is of primary interest since this is one target unit for the census and for future production of official statistics where a clear picture of the number and types of households is vital. The survey data will be compared with the register data to get estimates of net and gross error. For a categorical register variable Y, both overall gross error and gross errors of the respective categories can be computed. It is derived by comparing the evaluation sample data and the register data in a two-way $J * J$ table counting the objects with respect to the values of Y. If we put the observations from the evaluation survey in rows, and those of the register in columns, the estimated proportion of the overall gross error is the proportion of misclassified objects in the register (i.e. the off-diagonal proportion in the two-way table). The proportion of gross error in a category is given by $GE_j = (N_{\bullet j} + N_{j\bullet} - 2N_{jj})/N_{\bullet j}$, where N_{jj} is the number of objects of category j in both the evaluation and register. $N_{\bullet j}$ and $N_{j\bullet}$ indicate summation over rows and columns respectively. Gross errors reflect a particular aspect of reliability that does not concern the quality of estimates from the register data, but which are of interest for reliability in general. The size of the gross error of the register variables is relevant when data are used in different kinds of analysis, for example as a basis for forecasts and projections or to study changes with flow models. Large gross errors may destroy the possibility of using the material for analysis and also be a sign of poor measurement quality. As users of the registers, we should take care if aggregates are calculated for subpopulations or when we use the information to construct frames and stratification in sample surveys. The net error proportion is given by $NE_j = (N_{\bullet j} - N_{j\bullet})/N_{\bullet j}$. It gives information about the error in a category when the register is used for aggregates.

We illustrate by looking at the type of household category *Cohabitant* (non-marital partnership). The estimates are taken from a cross-comparison of the TPR and questions embedded in the Swedish Labour Force Survey (LFS), which were used to derive a register-based variable reflecting cohabitation. Cohabitation derived from the TPR is defined as a man and woman with less than 15 years age difference living at the same address.

Table 2: Cross-classification of cohabitants according to LFS and TPR

	TPR	
LFS	*Cohabitant*	*Not Cohabitant*
Cohabitant	19.7 %	3.3 %
Not Cohabitant	3.5 %	73.4 %

The gross error estimate is 3.5+3.3=6.8 percent and the net error is 3.5-3.3=0.2 percent. The latter indicates that counting the number of cohabitants from the TPR would not give

much error in the estimate. The gross error on the other hand can have an impact if cohabitants are a domain of study in an analysis of another variable.

4.2.2 Missing data

The inflow of data for the evaluation study was 65 percent. The goal was to reach a response rate of at least 60 percent, and at least 40 percent in each stratum. Both goals were met. These are realistic goal but still a large amount of nonresponse and it is thus important to analyze the response pattern and to have a strategy for how to handle missing data.

Results by Englund and Nilsson (2012) show that the nonresponse is higher among those who are without a dwelling id-key in the TPR, particularly in the Stockholm region. Nonresponse is also higher among people aged 35-44, those with no or low taxable income and individuals born outside the Nordic countries. Renting or owning the dwelling does not seem to matter in terms of nonresponse but, as can be expected, nonresponse is higher among individuals who are registered in dwellings that have three or more families registered. This is probably an indication of a higher amount of incorrect information in the register for those records.

4.3 Lessons learned

It is still too early to present conclusive results from the study. However, there are some observations that can be made so far.

The first experience is the difficulty to construct a web/paper question designed to collect information about the household status. We used a template from a previous census evaluation, but it still took quite some time to test and decide on the final design. This was not only because of slightly different goals compared to previous evaluation or technical difficulties in the web questionnaire environment. Collecting information in order to identify type of household and its members is most likely simpler by an interview (a comparative study has not yet been done).

Secondly, and not unexpectedly, we observe that the nonresponse is higher in the groups where the Tax Agency also had difficulties collecting information and where we suspect errors in the register. The stratification variable, *Number of families in dwelling according to TPR*, has a category with three or more families per dwelling. This is very unusual and any number above zero is an indication of registration errors. High nonresponse could be an effect of not reaching the individuals since they do not live where they are registered. A similar explanation can be brought up for the low response rate (57 percent) among those missing the dwelling id-key. However, it does not explain the lower response rate in the youngest age strata. That observation follows the same nonresponse pattern as in other Swedish surveys. One theory is that, since young adults are more mobile, the register might not be updated fast enough and they are therefore harder to reach. Another theory is that their propensity to respond is low, including low propensity to report to the Tax Agency when moving. This could be studied by analyzing variables registering events of change in the TPR together with the evaluation survey. Hence, there is a suspicion of lower register quality in the younger adult population but presently we cannot see any clear connections.

Although anticipated, we must stress that the analysis and reconciliation of the household composition demands plenty of resources. For the type of household variable, the various

combinations of discrepancies that might exist between the register information and the survey mean that the work requires a lot of manual handling. We also have to make judgments about which source to trust, and when.

5. Final Remarks

The evaluation survey of the Dwelling Register and the household statistics is going to give Statistics Sweden important information about some of the errors and the quality of the register-based statistics. It will provide information to at least describe the strengths and the weaknesses of the data, and hopefully it can also be used to adjust the census statistics if it is considered necessary. The study will be valid for the time frame close to the census date. However, since evaluations are expensive, this study is likely to be referred to for a longer time than that. As far as the Dwelling Register is concerned, it is probably low risk in doing so. It is completely new and it is likely that the register will improve as time goes by, and the administrative routines improve.

There are aspects that are not covered by this evaluation survey. As far as the census is concerned, it does not cover the quality of other register based variables such as education, employment etc. For these we will have to rely on older studies and descriptive measures of quality. Furthermore, to evaluate the quality of the housing statistics a different approach is needed. This is because much of the information has to be collected from property owners and municipalities, rather than from the population of individuals. A planned study will focus on the dwelling population and the statistics on dwellings that do not have a registered individual.

Since Statistics Sweden plans to develop a system for future household statistics that is register-based, and since some core components will be taken from the census production environment using registers over and over again, a plan to continuously monitor quality in the registers should be developed. This can be done in several ways, e.g. embedded questions in regular surveys, quality studies by observations and comparisons, monitoring event variables, being proactive in the collaboration with the authorities that create the registers, evaluation surveys like the one described here, etc. To put together a cost-efficient set of methods and tools to do this must be a priority in the years to come. This requires both theoretical and empirical development work, but we believe it is necessary in order to keep the reputation of being a producer of trustworthy official statistics.

References

Axelson M., Hedlin D., Holmberg A., and Jansson I. (2010). Methodology in the Swedish register-based census. Paper presented at the 2010 International Methodology Symposium, Statistics Canada, Ottawa, October 26-29.

Eltinge, J. L. (2010). Accounting for temporal effects, sampling variability, incomplete data and reporting error in the integration of consumer expenditure data from survey and administrative-record sources. Paper presented at the 2010 International Methodology Symposium, Ottawa, Canada.

Englund, J. and Nilsson K. (2012). Bortfallsanalys på evalveringen av Census 2011.
Unpublished project report, Statistics Sweden. (In Swedish)

Hedlin D., Holmberg A., Jansson I., and Lorenc B. (2011a). The first fully register-based census in Sweden. Paper presented at the 2011 Joint Statistical Meetings, Miami Beach, July 30– August 4 2011.

Hedlin D., Holmberg A. and Jansson I. (2011b). Combining registers into a fully register-based census – some methodological issues. Paper presented at the NORC U.S. Census Bureau conference: Utilizing Administrative Data: Technical, Statistical and Research Issues October 27-28, 2011, Washington D.C.

Wallgren, A. and Wallgren, B. (2007). Register-based Statistics - Administrative Data for Statistical Purposes. New York: John Wiley.

Zhang, L.-C. (2011). A Unit-Error Theory for Register-Based Household Statistics. Journal of Official Statistics, 27, 415-432.

Zhang, L.-C. (2012). Topics of Statistical Theory for Register Based Statistics and Data Integration. Statistica Neerlandica, Vol 66, No 1, pp 41-63.