Statistiska centralbyrån
Statistics Sweden

Ingegerd Jansson

Issues and plans for the disclosure control of the Swedish Census 2011

Abstract

The Census 2011is the first census in Sweden since 1990 and it will for the first time be fully register-based. In addition, the system of registers now created will be the foundation of all future official statistics on housing and households. In order to provide a feasible method for disclosure limitation, several strategic decisions have to be taken. The paper discusses important issues and some possible methods for protecting data.

1 Background

1.1 Introduction

To protect data properly before it is disseminated is important to all NSI:s, for legislative and ethical reasons, and for the credibility of the agency. Statistics Sweden is no exception. Results from Census 2011 are to be delivered to Eurostat no later than March 2014. However, Statistics Sweden aims at publishing results nationally in the autumn of 2013. The work on finding a feasible solution for disclosure control was initiated in a first phase of the census SDC project in 2011, mainly consisting of a literature review. The resulting report lists feasible methods, reviews the work of other statistical agencies, and points to important considerations. Statistics Sweden is now launching the second phase of the project, at a point where strategic decisions have to be taken in order to decide how to proceed.

For the first time in Sweden, the census will be completely register-based. The system of registers that are now in place will not only deliver data for the census, but, more importantly, it will be the foundation for future official statistics on housing and household statistics. Thus finding a solution for statistical disclosure limitation in the census data is integrated with the more extensive question on how to protect housing and household statistics in the future.

This paper starts by briefly describing the premises for a register-based census, and then proceeds by listing important issues and considerations that has to be taken in order to find a satisfactory solution for the disclosure control of the Census 2011 data, as well as future official statistics on households and housing. Some possible methods for protection of tables are mentioned.

1.2 A register-based census

Census 2011 is the first fully register-based census in Sweden. Reference day is 31 December, thus time was allowed for the necessary registers to be properly updated during 2011. The last traditional census in Sweden

was conducted in 1990, with data being collected by a self-administered mail-out mail-back questionnaire. In 1995, the Swedish parliament took the decision that the next census should be completely register-based. For several reasons, among them political concerns of privacy, the necessary legal regulation was not in place until more than ten years later. Still, a register-based census was not feasible due to the lack of a link between the Total Population Register (TPR) and the Real Property Register (RPR). In 2007, the Swedish parliament passed a bill on the creation of a new Register of Dwellings (RD), including flats and single houses, which will be linked to both the TPR and the RPR.

The Census 2011 is obviously a very important task, but for Statistics Sweden an equally important advantage of building a complete system of registers on individuals and real property is the possibility to improve official statistics on households and housing. Statistics Sweden will be able to provide official statistics in areas where there has been considerable shortage so far. An up to date and coherent system of registers enhances the capability to produce statistics on demand with improved longitudinal quality at low cost, produce statistics on smaller domains and special populations, provide standardised register variables and populations that will improve the comparability between studies, and provide a better framework for sample surveys.

These new opportunities also raises privacy issues and put high demands on Statistics Sweden to properly protect disseminated data. Administrative registers are used extensively in Sweden, at municipal as well as state level. There is a history of political debating about privacy concerns regarding administrative data and statistical use of such data. The traditional census method of mailing out questionnaires was questioned by the public in 1990, and the general view at that time was that the use of administrative data was perceived as less intrusive than a mailed questionnaire. Today, the use of administrative registers is not questioned per se, but privacy concerns are of great importance. As a consequence of the previous privacy debate, Statistics Sweden is not allowed to store the census micro data for future use.

The system of statistical registers relies on three core registers: the Business Register (BR), the TPR, and the RPR. The core registers are linked to various subject matter registers such as registers of employment, occupation, education, and buildings. The system of statistical registers is dependent on the possibility to uniquely link information to objects. Comprehensive and unique identification numbers of persons and businesses have existed for a long time in Sweden. With the creation of the RD, a unique identification number has been given to all dwellings (flats and houses). This allows for a definition of households based on the registers. Further information on the Swedish register-based census and related methodological concerns can be found in Axelson et al (2010) and Hedlin et al (2011).

2 Important issues

2.1 Scope

There are two conceivable ways for Statistics Sweden to proceed when deciding on a proper method for disclosure control. We could start with a specific disclosure control solution for the Census 2011. This includes strategic decisions and choice of proper methods, as well as providing the technical means for carrying out risk assessment and protection, taking into account Swedish legislation and the requirements of Eurostat. Or we could aim for a more general solution that may apply to all types of published statistics on housing, households and individuals, as fixed tables or in a flexible on-line system. If we choose to focus on a specific solution for the census, working on a general solution still remains.

If Statistics Sweden decides to aim for a more general solution at this point, there are still specific issues for the census to be solved. A similar statement holds the other way around: if a specific census solution is prioritized, it is highly desirable that the results and lessons learned can be utilized in a future general solution.

2.2 Legal requirements

According to EU requirements, data should be protected by the national agencies, and national legislation applies (Regulation No 763/2008). The premise of the Swedish legislation is that all official statistics are confidential without exception and may only be published if it can be guaranteed that a disclosure would not cause harm or damage to an individual.

2.3 Strategic issues

There are several strategic issues that Statistics Sweden needs to consider before determining which methods to use for risk assessment and protection of data. An important issue is where the greatest risk is considered to be. Are small values, detailed geography, attribute or identity disclosure, group disclosure, risk of differentiation, or anything else considered to be the biggest threat? These decisions should be taken at a level of the organization where the formal responsibility for confidentiality and privacy issues lies, i.e. management. Once these strategic issues are settled, appropriate methods can be chosen. It is important for Statistics Sweden to reach consensus in order for the methods chosen to have an impact in the organization.

Office for National Statistics is an example of an agency that has done considerable work in order to arrive at a policy for the protection of census data (Longhurst et al 2009, Spicer and Tudor 2009). We are considering how to achieve something similar (only with limited time and resources) for the Census 2011, and in general how to work in a systematic way to assess risk scenarios in statistical products, using Census 2011 as a case study. Such a systematic way of working (a framework or check list) might in the end lead to a general statement or

policy on how Statistics Sweden views risk.

Another important consideration is what to view as most important to keep in the published data. Risk must be reduced to an acceptable level while at the same time not causing unacceptable reduction of the utility of the data. Particular concerns where disclosure control methods may significantly affect data should be discussed with important users of data so that disclosure risk can be balanced against their requirements. Consistency between tables, additivity within tables, and the effects that disclosure control might have on common analyses are important to discuss with users. In this context, it is also important to inform users that we are protecting data and why we do it.

Census tables will be published in the Swedish Statistical Databases (SSD), an on-line system with some flexibility accessed from the webpage of Statistics Sweden. From SSD, Eurostat will retrieve data, and any other users will have access to the census tables. However, the SSD does not currently have a function for protecting data. The choice is either to protect tables before the aggregated data is uploaded on to the system, or to add functionality in SSD so that data is protected when there is a request from a user. Here too the aspect of specificity to census or generality to future housing and household statistics applies. If SSD is redesigned for disclosure control, it is a definite advantage if the solution can be utilized in a future general solution. In addition, it has implications for other types of publication since the methods performed on tables based on the same data set that are published in any other way should not differ from the methods used in the on-line system, for consistency between tables and to avoid confusion among users but also to limit the possibility of disclosure through e.g. different suppression patterns, differencing etc.

3 Feasible methods

The strategic issues mentioned above need to be considered before a definite choice of method for data protection is made, but the report from the first phase of the census SDC project points to some possible methods. Finding a strategy and criteria for choosing a method, as well as making a choice, will be the task of the second phase of the project. Lessons can be learned from strategies adopted at other agencies, described for example in Forbes et al (2009), Longhurst et al (2009), Spicer and Tudor (2009), and Camden et al (2007).

Municipality will be the lowest level of geography, occurring in only a few of the hypercubes. There are 290 municipalities in Sweden and there were 9 482 855 persons registered in the TPR as of 31 December 2011. Clearly, tables will be sparse with many zeros or low values. The final data set does not yet exist, but for phase two of the census SDC project, enough data will be available in order to perform tests of protection methods and choices of parameters.

For a general solution that will hold for the future housing and household statistics, a pre-tabular method is probably to be preferred. New data will be collected from the system of registers at given time points and tables will be presented in SSD, as fixed tables, and on user requests. To ensure consistency between different forms of output, and to reduce work load between data collections and complexity of IT-systems, it seems most efficient to protect micro data once for each data collection instead of continuously protecting tables as they are uploaded or requested from SSD. Targeted record swapping is an option (see e.g. Shlomo 2007, 2009). Record swapping in a census context usually means that geographical variables are swapped between random pairs of households, controlling for some household characteristics in order to minimize bias. Targeting means that households most at risk for disclosure are more likely to be swapped. Swapping would give consistent and additive tables. The method can be supplemented with additional checks and specific rules if necessary.

The ABS cell perturbation method could also be considered, designed to provide protection for flexible tables in particular where differentiation is be a considerable risk (Fraser and Wooton 2005, Leaver 2009). The method assigns a value to each micro data record and these values are aggregated to cells. The cell-level key and a look-up table is used to determine the amount of perturbation applied to each cell. Each cell will have the same perturbation independent of the table where it appears. Applying the method will give consistent tables, but they are not necessarily additive. Additivity can be restored, but then consistency is affected.

Targeted record swapping is feasible also for a specific census solution, with the additional benefit of being a case study for the general solution. However, it has also been discussed to use post tabular methods for the census, as this is a one-time event and data will only be published through SSD. One option is to use random rounding of small values, possibly together with additional rules like a minimum average cell value or a threshold for the number of households in a municipality. An important choice is to decide when disclosure protection is applied (before uploading in SSD or as tables are requested). With a specific solution, the amount of resources and work put into the disclosure protection should be reasonable.

Households will be created based on register information on individuals, i. e. marriages, children – parent relation, address, and apartment number. Additional approximate rules are necessary to separate families from people sharing apartment but not household. There is likely to be missing data due to incomplete addresses and missing apartment numbers. This means that we will have seemingly “empty” apartments and “homeless” people due to missing register data.

The problem will probably be addressed with imputation or weighting. The data on households and how households are created will be evaluated by a sample survey (data collection in March and April 2012). The

evaluation will give us a hint of the amount and types of measurement errors. There is no reason to protect imputed values (or for that matter values affected by other systematic errors that create uncertainty in the data, if they can be measured) since that would be information lost in vain. With targeted record swapping, imputed records can be excluded from swapping. The look-up table of the ABS cell perturbation method can be specified so that imputed values are not perturbed. We hope to be able to put more effort into relating disclosure control to the total error of register data.

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.

Camden, M., Cowie, P. and Henley, L. (2009). Census Tables: Utility and Safety via a Cell Threshold. Work session on statistical data confidentiality Manchester 17-19 December 2007. Eurostat Methodologies and Working papers. European Commission 2009.

Forbes, A., Naylor, J., Leaver, V., Gare, M., Hawkes, T., and Camden, M. (2009). Confidentiality Plans for the 2011 Censuses in the UK, Australia and NZ: a Comparison. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Bilbao, 2-4 December 2009.

Fraser, B. and Wooton, J. (2005). A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing. Paper presented to the Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, Switzerland, 9-11 November 2005.

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

Longhurst, J, Tromans, N, Young C, and Miller C. (2009). Statistical Disclosure Control for the 2011 UK Census. Work session on statistical data confidentiality Manchester 17-19 December 2007. Eurostat Methodologies and Working papers. European Commission 2009.

REGULATION (EC) No 763/2008 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 9 July 2008 on population and housing censuses. OJ L 218, 13.8.2008.

Leaver, V. (2009). Implementing a Method for Automatically Protecting User-Defined Census Tables. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Bilbao, 2-4 December 2009.

Shlomo, N. (2009). Statistical Disclosure Control for European Census Dissemination. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Bilbao, 2-4 December 2009.

Shlomo N. (2007). Statistical Disclosure Control Methods for Census Frequency Tables. International Statistical Review, 75:199-217.

Spicer, K., and Tudor, C. (2009). Balancing Risk and Utility – Statistical Disclosure Control for the 2011 UK Census. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Bilbao, 2-4 December 2009.