

WP. 28
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (v): Statistical disclosure control methods for the next census round

**CONFIDENTIALITY PLANS FOR THE 2011 CENSUSES IN THE
UNITED KINGDOM, AUSTRALIA AND NEW ZEALAND: A COMPARISON**

Invited Paper

Prepared by Angela Forbes and Jane Naylor (Office for National Statistics, United Kingdom),
Victoria Leaver and Melissa Gare (Australian Bureau of Statistics, Australia),
Tim Hawkes and Mike Camden (Statistics New Zealand, New Zealand)

Confidentiality plans for the 2011 censuses in the United Kingdom, Australia and New Zealand: a comparison

Angela Forbes^{*}, Jane Naylor^{*}, Victoria Leaver^{**}, Melissa Gare^{**}, Tim Hawkes^{***}, Mike Camden^{***}.

^{*}Office for National Statistics, Segensworth Rd, Titchfield, Fareham, PO15 5RR England, angela.forbes@ons.gsi.gov.uk, jane.naylor@ons.gsi.gov.uk

^{**}Australian Bureau of Statistics, Locked Bag 10, Belconnen ACT 2616 Australia, victoria.leaver@abs.gov.au, m.gare@abs.gov.au

^{***}Statistics New Zealand, Box 2922 Wellington, New Zealand, tim.hawkes@stats.govt.nz, mike.camden@stats.govt.nz

Abstract: Several National Statistical Institutes (NSIs) are preparing for censuses in 2011 and an important part of this includes the development of statistical disclosure controls (SDC) to protect the census outputs. The development of a SDC policy for census is a large undertaking and typically involves evaluating past methods, consulting with users, considering census outputs and data access methods and determining an overall confidentiality strategy. This paper will compare three NSIs; Office for National Statistics (ONS) for England and Wales, Australian Bureau of Statistics (ABS) and Statistics New Zealand (StatsNZ), as they plan the disclosure controls to protect their census data. Differences in their goals, legislation and attitudes towards data utility and risk, lead to different decision making paths and processes, and ultimately to different SDC strategies being proposed. Examples from each NSI will be used to highlight the issues considered and the decision processes used. By comparing these international differences we can gain insights into the factors which drive the selection of various SDC methods.

1. Introduction

The United Kingdom, Australia and New Zealand all have long histories of carrying out respondent-based censuses, and each have plans for a Census of Population and Dwellings in 2011. Publishing census data carries the risk that individuals could be identified and private information about them could be released therefore NSIs need to protect the confidentiality of census respondents. The production and use of official statistics depends on the cooperation and trust of citizens and this trust cannot be maintained unless the privacy of individuals' information is protected otherwise confidence in the NSI falls and response rates are also likely to fall. Each NSI also has legal obligations to protect respondents' information and census data, and complying with these regulations is the first requirement of any disclosure control strategy.

The UK Statistics and Registration Services Act (2007) prescribes a Code of Practice for Official Statistics which '*ensures that official statistics do not reveal the identity of an individual or organisation, or any private information relating to them, taking into account other relevant sources of information*'. The Statistics and Registration

Services Act also stipulates a fine and/or imprisonment for persons contravening the act and disclosing personal information.

The Australian Census and Statistics Act (1905) states that the ABS shall not release statistics *‘in a manner that is likely to enable the identification of a particular person or organisation’*. To meet this requirement, the ABS has developed a number of policies one of which specifies that table cells containing very small counts should not (generally) be released. In addition, the ABS should not release groups of tables that allow users to derive the true values of cells with very small counts.

In New Zealand the Statistics Act (1975) states *‘All statistical information published ... shall be arranged in such a manner as to prevent any particulars published from being identifiable by any person’*. To apply this legalisation StatsNZ has created a Confidentiality Standard for Census (Statistics New Zealand 2009), which specifies *‘cells with only one or two contributors are considered to be at high risk of disclosing individual information. (They) will be systematically modified to introduce uncertainty about the true value.’*

In addition to meeting these legal requirements to protect individuals’ information, each NSI aims to maximise data utility while preserving respondent confidentiality. However, histories and attitudes differ between the countries and different perspectives towards risk and utility (together with other influencing factors) leads to different statistical disclosure control (SDC) methods being proposed by each NSI. This paper explores the differences between the ONS, ABS and StatsNZ. The factors influencing decisions around confidentiality are considered in section 2. The different decision making processes adopted by the three NSIs are described in section 3, along with the proposed SDC strategies for the 2011 Census in section 4. This paper focuses on the protection of census tables rather than other products such as microdata samples.

2. Influencing factors

In developing disclosure control policies for their 2011 Censuses, each NSI started by evaluating the methods used in their previous census. An important part of this involved considering the feedback provided from users and the impact of SDC on the utility of the census data. The perceived level of risk posed by census tables was considered along with the feasibility of various approaches in combination with output strategies, geographic classifications and IT requirements.

2.1 Lessons learnt from previous censuses

In the previous census in 2001, the ONS used record swapping and small cell adjustment as the main methods of disclosure control for tabular outputs. Record

swapping created uncertainty in the census data by exchanging geographic information between randomly selected pairs of households. Small cell adjustment was then applied to add more protection for the small cells in tables. *The Conduct of the 2011 Censuses in the UK, Statement of Agreement of the National Statistician and the Registrars General for Scotland and Northern Ireland* (2008). Scotland and Northern Ireland run concurrent censuses with England and Wales, but in 2001 the disclosure control rules varied across the countries. This inconsistency between the UK NSIs caused problems for users wanting to compare the UK outputs. To address this issue for 2011, the Registrars General of Scotland, Northern Ireland, England and Wales agreed to work together and aim for a common UK SDC methodology. The ONS lead this work by considering alternative SDC methods for 2011.

For the previous Census of Population and Housing in 2006, the ABS developed and implemented a new confidentiality method. This method perturbed the census data by applying a permanent numeric value to each unit record in the census data set called a "record key". When a table was created, the record keys for each unit record in a cell were combined to give a key for the cell. This cell-level key was then used to determine the perturbation that is applied to the cell via a fixed look-up table. For more information about the method see Fraser & Wooton (2005). This method was designed to allow much more flexibility so that users would be able to define their own tables and each table would be protected consistently. This new method was a success in 2006 so the ABS is planning to retain and improve it for their 2011 Census.

StatsNZ also plans to adopt a minimal-change approach for 2011. This will be achieved by starting with the methods used in the 2006 Census and improving them by targeting utility and safety, and simplifying the rules. The original 2006 SDC rules included the suppression of tables where:

- the geographic classification was the most detailed (i.e. 'meshblocks', averaging 100 people)
- tables are detailed and include the income classification
- the mean cell size is low (this is defined as population for a geography divided by number of cells in the table)

In addition, all published counts were randomly rounded to base 3.

After the 2006 Census, StatsNZ responded to user needs for information that matched previous censuses and in 2007 StatsNZ began releasing tables with lower mean cell sizes, when these tables had small counts suppressed. Detailed travel to work data was also released to local government users under a licence agreement. The 2007 confidentiality rules are available online (StatsNZ, 2007)

Consistent SDC methods between censuses are beneficial because users develop expectations about which data is available and how it is presented. Consistency needs to be balanced against the potential improvements which can be made from updating

or changing the SDC methods used. ABS and StatsNZ conduct censuses every 5 years whereas the ONS holds censuses every 10 years. The greater time between censuses means there are likely to be larger changes as technology and best practice can shift considerably in 10 years. This additional time between censuses also allows the ONS a longer development period for all methodologies including SDC. The ABS and StatsNZ evaluated their 2006 approaches to SDC and for 2011 they have adopted minimal change strategies. After considering their 2001 SDC approach, the ONS selected to evaluate a range of different SDC options for 2011.

2.2 Risk Thresholds

SDC methods aim to reduce the risk of a statistical disclosure to an ‘acceptable level’. The level of risk deemed acceptable varies across the NSIs. Small cells appearing in census tables can lead to actual and perceived disclosure risks. Actual risks include identity disclosure which occurs if an individual is able to be identified from the data. Identity disclosure can include a person finding their own response in a table of census data. Attribute disclosure is another risk which can occur if an intruder is able to identify an individual and then learn new information about them from the data. Perceived disclosure risk refers to the public’s feelings of risk associated with different outputs. Identity and attribute disclosures can also occur from larger cells counts but concerns around public perception have meant that traditionally the focus has been to manage disclosure risk by removing or protecting small cells. In their previous censuses the ABS, StatsNZ and the ONS all focused on removing small cells in census tables which (along with some modification to larger cell values), guarded against identity and attribute disclosures. In 2011 StatsNZ and ABS intend to continue with this approach, but the ONS has shifted its focus. For 2011 the Registrars General from the UK NSIs have agreed that small counts (e.g. 0’s, 1’s and 2’s) can be included in publicly disseminated census tables provided that sufficient uncertainty as to whether a small cell is a true value, has been systematically created. In this policy the Registrars General identified attribute disclosure as the key disclosure risk for the 2011 Census. *The Conduct of the 2011 Censuses in the UK, Statement of Agreement of the National Statistician and the Registrars General for Scotland and Northern Ireland.* (2008).

Another disclosure risk is called ‘differencing’ and this can occur if users are able to obtain several tables for similar populations and then take the difference between them. Differencing can reveal the counts and information for much smaller sub-population groups, than the NSI intended to release. The perturbation method used by the ABS in 2006 was particularly effective at protecting against differencing (and other multiple table SDC risks) because the cell keys enabled the same cell to be protected consistently across different tables. Each of the NSIs consider differencing an important risk for 2011, and the proposed methods all create uncertainty across the dataset so an intruder would not be able to establish an exact count through differencing.

The prevailing public attitudes towards risk, the apparent sensitivity of census information and the political climate, all impact the risk appetite of NSIs and therefore decisions made around disclosure control. In the UK the Registrars General indicated a lowering of the risk threshold between 2001 and 2011 by focusing on attribute rather than the identity disclosure. Stats NZ also adopted a more lenient approach by adjusting their 2006 confidentiality rules in 2007 and they are also planning to simplify their SDC rules further for 2011. The ABS attitudes towards risk, as reflected in the confidentiality methods applied, have remained relatively consistent with their previous census.

2.3 Utility

Protecting against disclosure risk often comes at the cost of reduced data utility for users. It is therefore important to take into account how different users use the census data and balance these requirements against the disclosure risk. If data is not suitable for publication in census tables certain users may be able to access data through licensing arrangements.

After their last census the ONS gathered feedback from users and found several concerns with SDC related to:

- the lack of harmonisation of the SDC rules across the UK
- the late change to the SDC rules (where small cell adjustment was introduced in addition to record swapping)
- the impact of small cell adjustment on analyses and on user created geographies
- the delays in producing commissioned tables (these resulted from the need to manually check tables for disclosure risk)

For several years census data users in the UK have indicated a strong preference for consistent and additive tables. In 2001, small cell adjustment achieved table additivity (where rows and columns add to totals) by re-calculating the totals and sub-totals using the small cell adjusted data. But unfortunately these adjusted totals were not always consistent between different tables, e.g. the same cell could have different values in different tables. User consultation established that the majority of users were willing to trade-off some level of data quality in order to receive additive and consistent tables. To engage with users, documents explaining the SDC decision making process and the criteria used to evaluate SDC options, were provided on the ONS website (www.ons.gov.uk). Many user consultation events, which discussed SDC and possible impacts on outputs, were held throughout England and Wales with interested users. This approach ensured the ONS had a good understanding of user needs and users were able to have their preferences considered in the selection of the SDC strategy for the 2011 Census.

The ABS's consultation with their users found there was a strong and increasing demand for tables containing more detail. The users of ABS data were interested in options to create user-defined tables. In particular, users wanted to be able to define their own geographic areas for output. In response to this demand, the ABS developed two web-based products which were designed to be easy to use, and to provide a flexible range of data options. Users could create tables, maps and graphs using many combinations of data, and they could define their own geographic regions and custom groupings from a range of variables. One system: CData Online, is freely available to all Australian users CData Online allows users to define their own tables from several underlying topic-based datacubes. Another system: Census TableBuilder, requires users to be registered with the ABS, and this system allows users to create more detailed tables from most of the census variables. Several users expressed interest in obtaining more information about the ABS's SDC method. In particular they wanted to know how it works and the impact it would have on their analyses. Sophisticated users wanted to be able to adjust their analysis for the impact of the disclosure control. The ABS is working on developing measures of information and utility loss which could be incorporated into an information paper on the census SDC method.

Following the 2006 Census, StatsNZ met with some of their most intensive data users; from local and central government and universities; and the key findings from these meetings are summarised below. Users accepted StatsNZ's use of random rounding and the non-additivity that results from this method. The users supported the move to release larger counts in sparse tables and they were extremely keen for StatsNZ to recognise them as professional users who could be licensed to use more detailed tables. The users were somewhat ambivalent about StatsNZ's suggestions for new output-designed geographic classification as their historical series (using the existing geographies), are important for their analysis of trends. StatsNZ also consulted with internal users; including staff who respond to specific customised client requests; and these users placed a high value on having simple and easy to understand SDC rules.

2.3 Feasibility

Proposed SDC approaches must be congruent with other parts of the census development process including the census geographies and output strategies. Disclosure control rules have to be applied in time for data to be released, so in almost all cases this means they need to be automated into data processing (or output) systems.

In the 2001 Census, the ONS introduced a new geographic classification designed specifically for output. Previous classifications had been based on enumeration which had led to some areas having much larger and smaller than average populations. Consistent output areas with minimum population sizes helped

standardize the published data and also aided disclosure control. For 2011, only minimal changes will be made to the output areas used in 2001 which will also help across time comparisons. In addition, the ONS intends to increase the use of flexible tables in 2011, that influenced which SDC strategies were considered..

Geography was also an important consideration in developing the ABS's SDC strategy for 2006. The ABS wanted to allow users to build their own geographic regions from small areas. There was an increasing demand for more detailed data and for the flexibility to allow users to create user defined areas. These requirements remain the same for 2011.

The ability to automate their selected SDC method into existing software and systems was a significant consideration for StatsNZ, in both 2006 and 2011. It was also important for the SDC method to be simple in that it was easy to use and easy explain to users.

The SDC methods need to integrate into existing systems within the development time frame available. This, along with the lessons learnt from the previous census, the user preferences and the prevailing attitudes towards risk influence which SDC methods are short-listed and eventually selected by each NSI. If user feedback is generally positive and risk and utility thresholds relatively constant, only minor changes may need to be made to existing SDC approaches (e.g. ABS and StatsNZ). Alternatively if several influencing factors have changed, the process of evaluating potential SDC options is likely to be more intensive (e.g. ONS).

3. Decision Making Process

This section describes how the NSIs have adopted different approaches to select SDC strategies for their 2011 censuses.

To address the harmonisation issues which occurred in their last censuses, the UK NSIs agreed to work together to introduce a consistent strategy for 2011. The statement from the Registrars Generals discussed creating sufficient uncertainty in the census data and indicated the focus should be on protecting against attribute disclosures. This directive enabled the UK SDC working group (made up of representatives from each of the UK NSIs) to consider both pre-tabular (applied to the microdata) and post-tabular (applied to tables) SDC methods. The ONS conducted a filtering process where a range of potential SDC methods were evaluated against an initial set of qualitative criteria and then short-listed options were assessed against quantitative measures, and finally the options were ranked by the UK SDC working group. The qualitative criteria focused on additivity and consistency; user acceptability; the ability to protect against differencing; the feasibility of implementation; the ability to protect microdata; and the methods ability to be understood and accounted for in analyses. Each method was scored

against the qualitative criteria and three methods; record swapping, over-imputation and a method similar to ABS's cell perturbation, were short-listed for further investigation (Longhurst, Tromans and Young).

A sample of 2001 Census data was used to investigate each of the shortlisted methods. Risk-utility measures including; the percentage of small cells not perturbed, distance metrics, changes in variance and changes in the order of rankings, were compared on confidentialised output for a range of different tables. The UK SDC working group used the results from the quantitative analysis, along with their qualitative criteria, to agree upon a recommended method for 2011. This evaluation process was described in a report that was reviewed by SDC academics. The evaluation report will be available on the ONS website (www.statistics.gov.uk).

The decision making process for disclosure control at the ABS involved determining that the cell perturbation method used in the previous census, was suitable for use again in 2011. User feedback helped confirm that the method was appropriate. Once this decision was made further research was undertaken to refine the parameters of the method. The ABS went on to investigate how to incorporate the method into web-based table-building products, and the ABS's internal systems for creating and publishing census tables.

Similarly StatsNZ focused on making improvements to the confidentiality rules used in their previous census. The decision to adopt a minimal change approach was driven by StatsNZ's desire to maintain a relatively consistent set of confidentiality rules for users across the census years. Possible improvements to the existing rules were investigated and assessed across three dimensions: utility, safety and simplicity. Measures of safety and utility were calculated using a sample of the 2006 Census data. Measures of safety included the proportion of 0s and 1s in a table, the skewness of the distribution of cell counts, and the coefficient of variation of cell values. The proportion of 0s was used as a measure of safety in earlier SDC investigations (Camden, Cowie and Henley, 2007). Measures of utility included the number and proportion of suppressed cells and the proportion of people (or households, dwellings or families) in those suppressed cells. The simplicity of the rules was measured subjectively.

The three NSIs used different methods to decide on a SDC approach for 2011. The ONS received considerable user feedback from their last census and this together with a reduction in the perceived level of risk, led them to evaluate a wide range of SDC options for 2011. This meant the decision making process was longer and more detailed than that used by the other NSIs. The ABS had positive feedback from their 2006 Census, and with relatively consistent attitudes towards risk and utility, they selected to base the 2011 approach on the method they used in the previous census. StatsNZ responded to user feedback and refined their SDC policy shortly after their

last census, so their 2011 decision making process focused on simplifying their refined rules while maintaining data utility. Like the ONS StatNZ used quantitative measures of both risk and utility to inform these decisions.

4. Proposed SDC strategies and future work

The ONS, in conjunction with the other UK NSIs, have proposed targeted record swapping as the primary SDC method for protecting tabular outputs in their 2011 Census. Additional SDC protection will be provided through output and table design measures, and there are likely to be restrictions on releasing very detailed tables. Some of the more disclosive tables will only be available to approved users through licensing arrangements. The ONS also used a random record swapping SDC method in 2001, but in 2011 record swapping will be applied differently and it will be targeted towards areas and groups of respondents, which are considered a higher risk of disclosure. The ONS is currently working on how the targeted record swapping approach will be applied to the census data.

To protect their 2011 Census data the ABS intends to use a custom made cell perturbation approach. This method allocates each unit record a value and these are aggregated as records are used to make up cells. A lookup table is used to apply a certain level of perturbation to each cell in a table. The ABS is currently considering ways to extend this methodology to other data, so survey information can also be protected. Extensions of the method are likely to retain some features of the census approach, because the requirements for a survey table builder will be similar to the requirements for the census table builder. The ABS also predicts there may be scope to combine pre and post-tabular SDC methods e.g. by treating outliers in the microdata before the tables are created and confidentialised.

StatsNZ's proposed SDC strategy is based upon random rounding to a base of 3. This method provides the main protection for census tables and other disclosure control rules provide additional protection where random rounding is not sufficient e.g. for tables which contain a large proportion of small counts. The mean cell size rule states that the total unrounded subject population for a geographic area divided by the number of cells in the table, must be greater than two. If the mean cell size is less than or equal to two then all cells that have an unrounded count of five or less, are suppressed. Secondary suppression of other cells to prevent the original counts being derived does not occur but some unsafe scenarios can be predicted and these will be avoided by manual disclosure control. StatsNZ also intends to simplify the 2006 'meshblock rule' so that it better targets small counts. StatsNZ is currently investigating a proposal that any count greater than five, (for any combination of variables), can be published once it has been randomly rounded.

The three NSIs have selected three very different SDC methods to use in 2011. But within each NSI, the methods selected were quite similar to those each country used in their previous census. The ABS and StatsNZ took minimal change approaches so this similarity is expected, but the ONS evaluated a wide range of possible strategies before choosing record swapping. At this stage the SDC methods are proposals and it is possible they may change closer to the census day. Both the ONS and StatsNZ made significant amendments to their original SDC strategies, in their previous censuses.

5. Conclusion

NSIs have legal and ethical obligations to protect the confidentiality of their census respondents. As there is no prescribed way to meet these requirements, NSIs can choose the most appropriate methods which balance disclosure risk and data utility. The factors which influence the choice of SDC methods use in each country include; the lessons learnt from previous censuses, the attitudes and risk thresholds of the NSI, the impact on users and data utility; and the feasibility of approach in conjunction with other census systems and geographic classifications. Differences in these factors, and different decision making approaches, lead to the NSIs proposing different SDC methods for their 2011 Censuses.

The ONS received considerable feedback about the SDC methods they used in 2001, so in 2011 a wide range of possible methods were evaluated against a detailed set of qualitative and quantitative criteria. In line with the preference of the UK Registrars General, a pre-tabular SDC method (record swapping), was selected. The ABS developed a new perturbative SDC method for their 2006 Census. This method was successful in meeting user needs for consistent and flexible tables, so the ABS has decided to re-use this method for 2011. StatsNZ used random rounding and a series of other post-tabular methods, to protect their 2006 Census data. After the last census StatsNZ amended their confidentiality rules in response to user requests for historically consistent data. For 2011 StatsNZ intends to further simplify the 2006 confidentiality rules and continue to use a random rounding approach.

Although the three NSIs have proposed very different SDC approaches for 2011, there are several similarities in how they have approached the decision making process. In addition, the NSIs have all selected methods which are quite similar to those that they used in their previous census. By keeping SDC methods relatively consistent across censuses NSIs can fine tune and customise methods, and at a high level users can predict how SDC protection may effect the census tables they use.

References

Camden, M., Cowie, P. & Henley, L. (2007). Census tables: utility and safety via a cell threshold. Work session on statistical data confidentiality Manchester 17-19 December 2007. Eurostats European Commission 2009.

Fraser, B & 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.

Government Statistical Service (2009) *National Statistician's Guidance: Confidentiality of Official Statistics*. ONS.

Longhurst, J, Tromans, N and Young, C. (2007) *Statistical Disclosure Control for the 2011 UK Census*. ONS www.ons.gov.uk

Statistics NZ. (2009) *Confidentiality standard for Census*. (available on request.)
New Zealand. (1975) *Statistics Act*.

Statistics and Registration Service Act 2007, Her Majesty's Stationery Office and Queen's printer of Acts of Parliament.

Office for National Statistics(ONS), General Register Office for Scotland (GROS) and Northern Ireland Statistics and Research Agency (NISRA). *The Conduct of the 2011 Censuses in the UK, Statement of Agreement of the National Statistician and the Registrars General for Scotland and Northern Ireland*. (2008)
<http://www.ons.gov.uk/census/2011-census/produce-deliver-data/regrs-gen-agreement.pdf>