WP.38
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (vi) Censuses

EVALUATING IMPUTATIONS OF SEX AND AGE FOR SUBSTITUTES IN SUBSTITUTE HOUSEHOLDS

Supporting Paper

Prepared by Michael Ryan, Statistics New Zealand

Abstract

Unit non-response to a census is generally dealt with either by unit imputation within census processes (by creating substitutes); or, outside of census processes, by information gained through a census coverage survey (by estimating census undercount and overcount). For unit non-response from dwellings, substitute dwellings are created, with an imputed number of substitute individual forms, with sex and age also being imputed for these.

This paper describes using census coverage surveys to evaluate substitute dwellings created, and the substitute individuals created in them, in the New Zealand censuses of 2001 and 2006, focussing on evaluating the imputations of sex and age for substitute individual forms in substitute households. Census 2006 created about the right number of substitutes in substitute households, but Census 2001 created too many.

The distributions of sex were consistent between the respective censuses and their corresponding PES. However, the distributions of age were not fully consistent, with the census imputation procedures giving too many substitutes at older ages.

Keywords: age, sex, census, imputation, Post-enumeration Surveys, substitute forms, New Zealand.

I. INTRODUCTION

1. New Zealand is a small Pacific country, with an estimated resident population of 4.24 million at 30 September 2007. Census of Population and Dwellings are held every five years, and use both household (dwelling) forms and individual forms.

2. Population estimates are based on censuses, complemented by information on net census undercount from census coverage surveys*, and by information on residents outside the country at census, obtained from records of arrivals to and departures from the country (*Post-enumeration Surveys or PESs, as they are referred to in New Zealand). For each Census of Population and Dwellings, Statistics New Zealand develops initiatives to foster participation in the census. Despite such initiatives the unit non-response to census that manifests as substitute individual forms ("substitutes") has continued to rise; from 102,000 in 1996 to 107,000 in 2001; and to 133,000 in 2006. In 2006 substitutes were 3.2 percent of the estimated census population at census date. It is vital to understand the quality of the substitute

information created, as substitutes "are the main source of uncertainty in the census age-sex distribution" (Bycroft**,** 2007). Sex and age are the only variables imputed for substitutes in substitute households. This report only deals with substitute households / dwellings and the related substitute individuals created in them; it excludes substitute dwelling forms created to cover individual forms received but with no associated dwelling form.

3. The challenge of unit non-response to census is dealt with either by unit imputation within census processes; or, outside of census processes, by information gained through a census coverage survey. For unit non-response from dwellings, "substitute" dwellings are created; and a number of substitute individual forms are raised within each substitute dwelling; finally, sex and age are imputed for these substitutes. This paper describes how two PESs were used to: (a) evaluate the numbers of substitute dwellings created in the censuses of 2001 and 2006; (b) evaluate the numbers of substitute individual forms created in substitute dwellings in the censuses of 2001 and 2006; (c) evaluate the imputations of sex and age for substitute individual forms in substitute dwellings.

4. This brief paper is based on a review of substitute forms, whether raised for dwellings or for individuals, undertaken by Population Statistics staff and by Statistical Methods staff of Statistics New Zealand. The review is part of ongoing work into the quality of census processes and into the quality of estimates of the resident population. The review also had the task of providing a preliminary view of what methods should be used for handling substitutes in Census 2011. Substitute individual forms are also created within enumerated dwellings. The quality of information for these substitutes was not part of the review, nor of this report.

5. Section II describes the imputation procedures associated with the creation of substitute households. Section III outlines the PES methods that were used to estimate some characteristics of households and their members. Section IV compares the census detail on substitute dwellings and their substitute members with estimations from the PESs. Finally, section V discusses the assessment of the created details for substitutes, and looks ahead to Census 2011.

Terms

6. The term "dwellings" refers to the physical structures within which people live; and the term "households" refers to the groups of people living in those physical structures. In the New Zealand context, each private household is required to complete a dwelling form (DF); and the people in the dwelling on census night are required to complete individual forms (IFs), whether a member of the household or a visitor. Thus there is one dwelling form for every household.

7. The term "substitute household" refers to a combination of a substitute dwelling (form), together with the set of substitute individuals (forms) created within that substitute dwelling.

8. Meshblocks are the smallest geographic areas used in census operations.

II. IMPUTATION PROCEDURES

A. Imputation for substitute households

9. Speaking broadly, a household comprises the dwelling and the individuals living in it. Thus the creation of a substitute household involves the creation of a substitute dwelling form and the creation of substitute people to reside in the dwelling on census night.

10. Raising a substitute dwelling form is straightforward in that if it is determined that a dwelling was occupied on census night, and that no forms were received that related to that dwelling, then a substitute dwelling form is created. More specifically, to be a missing dwelling, the dwelling must be (1) listed on the dwelling frame; (2) classed as a private dwelling; (3) classed as occupied on census night; (4) have no forms received from it. Enumerator or processing errors can affect any of these four conditions.

11. For the censuses of 2001 and 2006 a Dwelling Control File (DCF) was used to control the formation of the census dwelling frame. The DCF starts out as an electronic file that is created before census, using information gathered during the previous census, and updated for new dwellings using building consents. The DCF was designed primarily to speed up processing. The DCF lists meshblock identifiers and line numbers, where a "line" (or record) and its associated line number corresponds to a dwelling believed to exist. Line numbers correspond to entries in the field-books used by enumerators. A perfect DCF would have one and only one line number for each dwelling; and with the right number of lines in each meshblock.

12. Fieldbooks have two main roles: they delineate where census forms are to be delivered to by enumerators; and they are used by enumerators to record the address and number of forms delivered to households.

13. When the field-books are received at the processing warehouse, the DCF is updated to agree with the returned fieldbooks. The information on occupied, unoccupied and new dwellings, and any unanticipated dwellings, is transferred to the DCF. New lines are added to cater for dwelling forms received that do not correspond to existing lines. Lines in the field-book that have not been used for a dwelling are deleted from the DCF. Once all fieldbooks have been returned and processed then the DCF is assessed for omissions and errors; and the DCF is finalised as the dwelling frame that accounts for all occupied and unoccupied dwellings.

14. In the balancing phase of census processing the DCF is used to "mark in" dwellings for which census forms have been returned. Marking in occurs after the dwelling identifiers are scanned and recognised as valid. Once all of the households have been marked in, the remaining occupied dwellings are considered outstanding. If a missing occupied dwelling is identified, a substitute DF and substitute IFs are created for it. Substitute households ("substitutes for missing households") are created throughout processing.

15. Improvements to the in-house process for reconciling the DCF with the fieldbooks used by enumerators gave good results in 2006. Some doubt still remains over how well enumerators are able to determine whether a dwelling is occupied ("lived in") or unoccupied ("not lived in"). Apartments and holiday homes are particularly problematic.

B. Imputation for the number of substitute individual forms per household

16. In 2001 and in 2006 a statistical approach was used, whereby the number of substitute individual forms to be raised for a missing household was taken from a donor household, randomly selected from within the immediate physical neighbourhood. Intuitively this is appealing, as it seems reasonable that missed households should be similar in size and composition to local households that are not missed by census enumeration. To be eligible as a donor household**,** a household had to have no more than six residents.

C. Imputation of sex and age for substitutes in substitute households

17. The information given here only applies to the imputation of sex and age for substitutes in substitute households.

18. In 2001 a joint age-sex distribution was imputed stochastically from the estimated resident population. Sex was imputed first, at 51% female to 49% male; and then age was imputed, conditional on imputed sex.

19. In 2006 a donor household methodology was used to impute age for substitutes in substitute households, whereby age was imputed from the donor household randomly selected from within the neighbourhood (the same donor household used to determine the number of occupants). Sex continued to be imputed stochastically, but was now imputed after age and independently of age. This was an oversight, as sex should have been imputed conditional on age.

20. Ideally both sex and age would be taken from the donor household. However information on the donor household comes from its dwelling form; and the dwelling form requests the ages of people who are in the dwelling on census night, but does not request the sex of such people. Thus in the census processing system only ages were available for the imputation of sex and age for substitutes. Consequently sex for substitutes in substitute households was imputed stochastically, independently of age (at 49% male, as applied in 2001).

21. In 2006, 96% of ages for substitutes in substitute households were imputed from age values in donor households. The remaining 4% of cases corresponds to when age was not recorded on the dwelling form of the donor household; for these cases age is imputed stochastically from a general distribution.

III. METHODS USED TO EVALUATE IMPUTATIONS

22. The main role of census coverage surveys is to measure overcount and undercount in census. After the 2006 census, the PESs of 2001 and 2006 were also used to evaluate the number of substitute dwellings created, the number of substitute individual forms created in substitute dwellings, and the imputations of sex and age for substitutes in substitute dwellings. The PES is the only source of estimates of the substitutes in substitute households that should have been created in census.

23. The New Zealand PES is a sample survey of people in private dwellings, carried out as independently as possible from census. Dwellings chosen for the PES sample are visited by PES interviewers shortly after census, with the 2006 survey being undertaken during the period from 21 March to 3 April 2006 (census was held 7 March). Personal details are collected, as are usual address, census night address, and any other addresses ("search addresses") where a person might have been included on any other census form. The matching of PES data to census data determines whether a person was counted, or not, in the census at each search address.

24. The presence of PES interviewers in the field may alter the pattern of late returns from what it would have been had PES interviewers not gone where they did. Carrying out the PES may have introduced a difference between what happens in areas that the PES surveyed, compared to what happens in areas the PES did not go to. In particular, the presence of PES interviewers may have prompted some households to return forms, which would then be classified as late returns*, but which would otherwise not have been returned at all (*census forms returned on or after 21 March 2006). This effect, if it did occur, is likely to have biased the estimation of number of substitute dwellings downwards.

25. The addresses of the PES dwellings selected for interviewing are the only ones that are used in estimating the number of substitute dwellings (other addresses supplied by respondents do not have a weight attached). In attempting to match PES respondents to census respondents, two variables are used to keep track of the possibilities; one records if the PES dwelling matches to a substitute dwelling; the other one records if a person is 'found' only in a substitute dwelling. The matching of PES respondents to census is not always successful. Where matching is successful, the match to census will be to either to a "real census form" or to a "substitute form".

26. For any address provided by a PES respondent, processing staff attempt to find a match in census. A match to a "real census form" means that a non-substitute individual form is found at that address, and it appears to correspond to the same person, the PES respondent.

27. However, another possibility is that an address provided by a PES respondent is matched to a census dwelling that is a substitute dwelling. The substitute dwelling form can relate to either a household for which no dwelling form was received but for which only loose individual forms were collected, or for which no forms were returned (i.e. a substitute household). For the first of these two possibilities, matching to a real IF is attempted. For the second, the match-code for the person at this address is recorded as a 'match to substitute'.

28. For substitutes in substitute households, the census number of occupants for a substitute DF is not likely to agree with the PES number of occupants, as the census number of occupants is imputed; nor will the census sex and age match the PES sex and age since they also have been imputed in the census. In this case, the PES values are taken as correct and used in the PES estimation of substitutes.

29. Thus for assessing details of substitutes in substitute households, we have two sets of data to compare: for the censuses of 2001 and 2006, the counts of substitute households, the counts of substitutes, and their sexes and ages - all of which are imputed; and from both post-enumeration surveys, the estimated numbers of substitute households, substitutes in substitute households, and their sexes and ages. The sample size of the PES (around 11,000 households in 2006) limits the detail of output available; in particular, age breakdowns were only available for the age groups 0-14 years, 15-29 years, 30-44 years and 45+ years.

30. Sample errors, corresponding to 95% confidence intervals, are used to assess whether the details of substitutes are reasonable. We use a simple test statistic **to** compare differences between census counts of substitutes (C) and PES estimates (P); the statistic is simply (C-P)/SE, where SE is the sample error that corresponds to the 95% confidence interval for P. Estimated sample errors are based on Jackknife methodology, using Generalised Regression (GREG) estimates, or based on calculated design effects. Most 2001 estimates of sample errors are borrowed from 2006.

IV. RESULTS

31. Simulation of the donor methodology used in the 2006 census confirmed that the methodology worked as expected, and did a good job of imputing households of expected size, and of imputing ages for substitutes in substitute households that are reasonable at national level.

32. In 2006, both the number of substitute households raised in census and the number of substitutes in substitute households are within PES sample errors, suggesting that substitute creation worked well in 2006 (Table 1).

Table 1 Census counts of, and PES estimates of substitute households and substitutes in substitute households, 2001 and 2006

	Substitute households		Substitute people in substitute households	
Census	**2001**	**2006**	**2001**	**2006**
Census count (C)	30,400	36,700	79,700	94,900
PES Estimate (P)	24,900	34,200	51,000	81,600
Census count - PES estimate	5,500	2,500	28,800	13,300
PES sample error (SE) (1)	7,000	6,900	15,800	15,800
(C-P)/SE (2)	0.79	0.36	1.82	0.84
Substitutes per household	-	-	2.62	2.59

(1) sample errors correspond to 95% confidence intervals;
(2) values outside the range of -1 to +1 indicate significant difference at the 95% confidence level.

33. However in 2001 the number of substitutes raised in substitute households in census is significantly higher than the PES estimate; and the number of households created in census was more than the PES estimated, although still within sample error bounds. These two results are both in the same direction, but with a stronger indication from the substitute people calculation (the estimates are likely to be positively correlated). The methodology for determining the number of IFs per DF has remained the same and produced very similar results in both censuses (Table 1). This suggests that probably more substitute people were raised in 2001 than should have been, and the cause was too many substitute households being raised.

34. Another possible interpretation of the 2001 results just mentioned is that the 2001 results simply reflect biases in the estimation of substitute details using the 2001 PES. However, we are unable to come up for reasons for why the 2001 PES would be biased so much in this regard and the 2006 PES was not. Our view is that such large biases are not plausible.

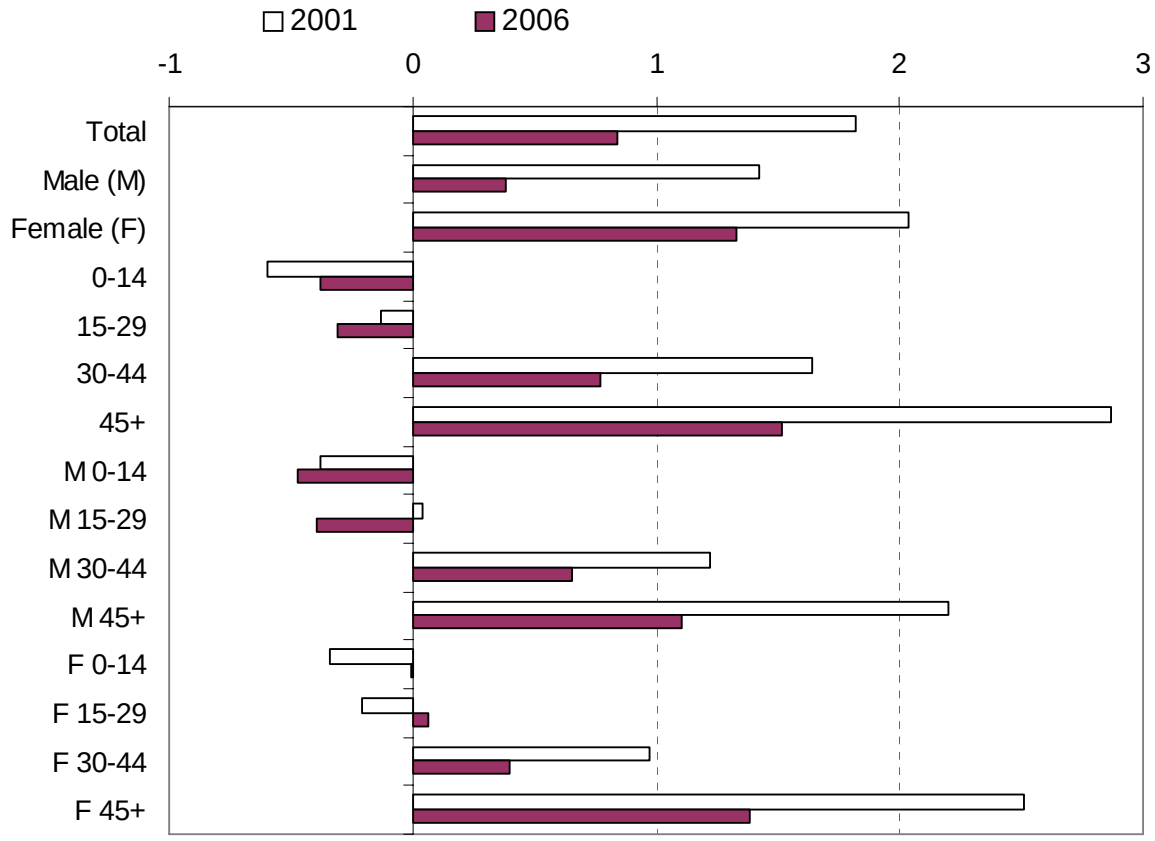
35. Figure 1 provides further detail, by presenting the statistics (C-P)/SE for various sex and age categories of substitutes in substitute households. The statistics should lie in the interval [-1,1], apart from 5% of the time. Neither census agrees fully with this pattern.

36. Census 2001 had too many substitutes of each sex. Too many male substitutes appear to have been created in the age groups 30-44 years and 45+ years; and too many female substitutes in the age group 45+ years.

37. Census 2006 appears to have created about the right number of male substitutes, but more female substitutes than should have been raised. The excess of female substitutes was most apparent in the age group 45+ years.

38. Further analysis done (only summarised here) showed that the sex proportions did not differ significantly between the censuses and the corresponding PESs. Unfortunately it was not possible to do a comparable analysis of age proportions. Analysis without sample errors showed that Census 2006 substitutes age group proportions were closer to those estimated from PES 2006, than were those of Census 2001 to PES 2001.

Figure 1 Substitutes in substitute households: Comparisons of census counts (C) and PES estimates (P) by means of the statistics (C-P)/SE



V. DISCUSSION

39. The total number of substitutes created appears to have been too high in 2001 and about right in 2006. This is consistent with the fact that more processing staff effort was able to be put into validating the creation of substitute households in 2006 than was possible in 2001. The most likely explanation is that too many substitute households were created in 2001. The distributions of sex were consistent between census and the corresponding PES. However, the distributions of age were not, with the census imputation procedures giving too many at older ages. The donor methodology used to impute age for substitutes in substitute dwellings worked as expected; and so the age bias noted appears to be a limitation of the methodology. In short, substitutes were of better overall quality in 2006 than in 2001.

40. While the causes of too many substitutes in 2001 remain unclear, possible sources of error include incorrect classification of occupied/unoccupied dwelling status by enumerators; and errors in the processing system reconciling the electronic DCF with the physical fieldbooks when finalising the dwelling frame. The statistical analysis of substitutes was hindered by being unplanned. This should be remedied for 2011.

41. The 2006 processing system improved on the 2001 system, and this probably contributed to the better results for 2006 substitutes. Options for further improvement for Census 2011 include:

improving checks between field work and the processing system;
including estimates of substitutes, dwellings and people, in the PES estimation system;
being prepared to adjust for an overcount of substitutes within the PES estimation methodology if necessary.

42. Previously census coverage surveys in New Zealand were seen simply as a vehicle to measure undercount and overcount in census. We now view such surveys as enabling a more comprehensive description of the quality of the census. Work is under way that may deliver a larger and better PES for Census 2011. This work includes assessing whether the methodology used by the Australian Bureau of Statistics in their 2006 PES is applicable in the New Zealand context (ABS, 2007).

References

Australian Bureau of Statistics (2007) *Measuring Net Undercount in the 2006 Population Census* (Information Paper).

Bycroft, C. (2007) "Challenges in Estimating Populations". *New Zealand Population Review 32 (2):21-47.*