Australian Bureau of Statistics

1352.0.55.092

Research Paper

Imputation and Estimation for a Thematic Form Census

www.abs.gov.au

New Issue

Research Paper

Imputation and Estimation for a Thematic Form Census

Philip A. Bell and Julian P. Whiting

Statistical Services Branch

Methodology Advisory Committee

16 November 2007, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) WED 22 APR 2009

ABS Catalogue no. 1352.0.55.092

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Philip Bell, Statistical Services Branch on Adelaide (08) 8237 7304 or email <statistical.services@abs.gov.au>.

CONTENTS

The role of the Methodology Advisory Committee (MAC) is to review and direct research into the collection, estimation, dissemination and analytical methodologies associated with ABS statistics. Papers presented to the MAC are often in the early stages of development, and therefore do not represent the considered views of the Australian Bureau of Statistics or the members of the Committee. Readers interested in the subsequent development of a research topic are encouraged to contact either the author or the Australian Bureau of Statistics.

APPENDIXES

IMPUTATION AND ESTIMATION FOR A THEMATIC FORM CENSUS

Philip A. Bell and Julian P. Whiting
Statistical Services Branch

ABSTRACT

The population Census provides a unique opportunity to obtain detailed information from the whole population of Australia in a way that supports tabulation for small geographic areas and fine classificatory items. This paper discusses a thematic form approach to the Census that aims to extend the number of items collected in the Census without increasing the respondent burden. The approach involves identifying a subset of the current Census items as core items (to be included on all forms), with the remaining questions plus a number of additional questions being arranged into 'themes'; in the simplest version each form contains a single theme.

This paper discusses design and estimation for a thematic form Census. It discusses the quality of estimates that could be produced at various levels under a weighting strategy, and the properties of these estimates. It then develops an imputation approach to producing estimates, in which values for items not collected on a particular form are imputed on the basis of information from similar dwellings providing the theme data. An investigation of the quality and properties of estimates under both a weighting and an imputation approach is presented.

The balanced imputation approach proposed in this paper appears to provide good outcomes in the thematic form context, with potential for application in a variety of other contexts.

1. SUMMARY

The population Census provides a unique opportunity to obtain detailed information from the whole population of Australia in a way that supports tabulation for small geographic areas and fine classificatory items. The number of items that can be collected in the Census is constrained by the need to avoid undue respondent burden, which would impact response rates and community acceptance of the Census. In the 2006 Census around 60 items were collected. There is a strong demand for good small area data of the type that can only be obtained by a large scale collection like a population Census.

The thematic form approach promises to extend the number of items collected in the Census without increasing the respondent burden. The approach involves identifying a subset of the current Census items as core items (to be included on all forms). The remaining questions plus a number of additional questions are then arranged into ‘themes’; in the simplest version each form contains a single theme.

The large size of the Census provides a huge sample for each theme question, allowing good small area estimates even though only a proportion of dwellings receive a theme question. This approach is a variation on the ‘long-form / short-form’ Census approach that has been used in a number of overseas agencies. Statistics Canada use a short Census form containing only core items for most dwellings, with a 20% sample receiving the long form containing additional detailed questions. Estimation in the short-form Census uses a weighting approach in which the long form data receives weights calibrated to a variety of short-form item totals for quite small geographic regions (Bankier, 2002).

The focus of the Census is on ‘small domain’ estimates – counts of persons in various quite small categories classified by geography and other items. Since the thematic form approach gives estimates (rather than counts) for these small domains, it is critical that the quality of estimates is sufficient to meet most user needs. The evaluation in this paper produces estimates by various approaches for a large ensemble of tables. It then evaluates the quality of estimates available from a thematic form Census by comparison to values from the full Census. A key requirement for a thematic form Census is that estimates for the theme items for various quite small domains will be of useful quality.

Section 2 of this paper discusses the design of a thematic form Census, and how various goals can be achieved by different arrangements of themes. Section 3 looks at the weighting approach to estimating for theme items, and discusses the properties of the resulting Census product. The accuracy of estimates that can be obtained under such a weighting approach is discussed.

Section 4 introduces the idea of using imputation to produce estimates. This would proceed by imputing values for theme items not collected on a particular form, on the basis of information from similar dwellings that provided the theme data. A simple 'hot-deck' implementation of this approach is described in which a similar unit is used as a donor for the missing items. The pros and cons of using individual persons or whole dwellings as donor units is discussed.

Unfortunately, the simple hot-deck will produce biased estimates, particularly for estimates by categories that do not play a major role in defining 'similarity' of donor units. Section 5 introduces the concept of a 'balanced' imputation, in which the imputation of theme items is controlled so that the imputed data will approximately reproduce a set of key tables. This allows estimates for these key tables to be of similar accuracy to those obtainable by weighting methods.

Methods for performing balanced imputation are discussed, and an approach is described that will be evaluated against other estimates. The approach was inspired by the balanced sampling method of Deville and Tillé (2004), but does not use that method (although Deville (2006) describes an application of balanced sampling in a random imputation setting). The approach developed in this paper more closely resembles the 'minimization' method proposed by Pocock and Simon (1975) in the context of randomised clinical trials.

Section 5 also describes the expected properties of estimates obtained under the proposed balanced imputation approach. These are compared to those of the weighted estimates and the simple hot-deck estimates described in earlier sections.

Section 6 describes an evaluation of the accuracy of the various estimators. The evaluation uses 2006 Census data after defining a subset of the items as theme items to be collected from only a subset of the dwellings. After choosing a sample of dwellings to provide the theme information, estimates under various approaches can be produced for a wide range of tables. The approaches are evaluated by summarising the differences between estimates and the true values across a large number of table cells, for a variety of tables categorised by combinations of theme items and core items and at different geographic levels.

Section 7 discusses the applicability of the balanced imputation approach described to other imputation contexts, and suggests further directions for development.

2. USING THEMATIC FORMS TO INCREASE CENSUS CONTENT

2.1 Increasing the content of a population Census

A fundamental breakthrough in survey statistics was the realisation that numbers of appropriate accuracy can be obtained by collecting data from only a sample of the population of interest. Applying this idea to a population Census, many items could perhaps be estimated with useful accuracy even if only collected on a subset of the forms.

A number of overseas agencies have moved to a long-form / short-form Census, in which a selected subsample of dwellings receive a full set of questions, while the majority receive a short form collecting a core set of items. This has the effect of reducing respondent burden for the majority of households and processing load for the agency. Statistics Canada found that a 20% sample of dwellings was sufficient to provide good estimates for long-form items at the required geographic output levels (Statistics Canada, 2002).

The thematic form approach is explored here with the primary objective of increasing the number of items that can be collected in the Census, by removing the constraint that all dwellings must receive the same set of questions. An alternative objective could be to reduce the number of questions any individual dwelling receives, thus reducing respondent burden and potentially improving Census acceptance.

2.2 Design of a thematic form Census

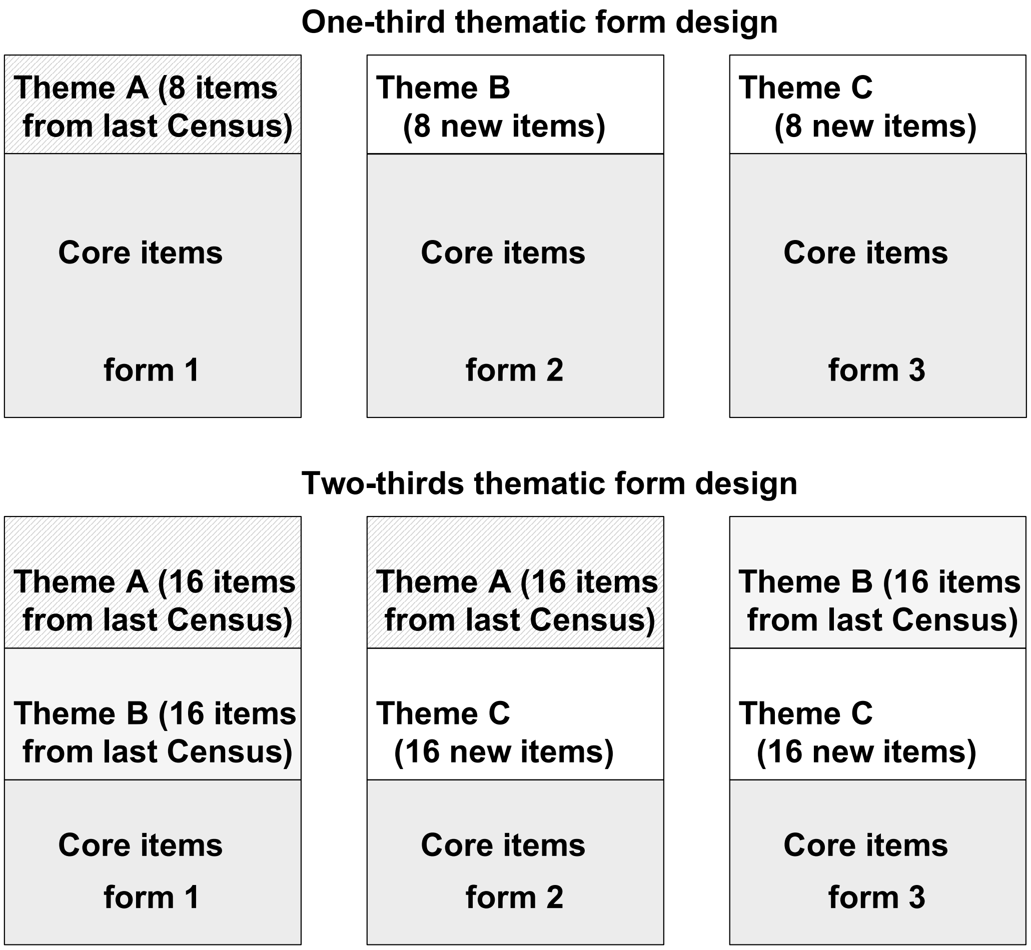
A first constraint in designing a set of thematic forms is the number of form types that can be used. This will be limited by costs and practical considerations. The most practical sample design appears to be interleaving the form types in the bundles of forms provided to Census collectors, so that each successive address on the collector's route receives a different form type. Under this design, having more form types will increase the variability of the number of forms of each type returned in an area. Having more form types will also increase questionnaire design, testing and printing costs, and may increase the likelihood of confusion among respondents and Census collectors. For this paper we will assume a design with three different questionnaires, each delivered to a third of dwellings.

A second constraint to be considered is what proportion of forms will contain a particular theme. With three forms, a particular theme can appear on a single form or on two out of three forms. Two designs have been proposed. The 'one-third' design would have three themes with a single theme on each form. This gives an expected sample size of one-third of the population. This design would retain the maximum number of current questions as core questions while accommodating a given number of new questions.

An alternative design is the 'two-thirds' design, in which each form gets two of the three themes, with a different theme omitted from each form. This design gives twice the expected sample size for each theme, markedly decreasing sampling error on estimates. The other advantage of this design is that every theme appears with every other theme on a third of the forms, allowing analysis of the interaction between theme items. A disadvantage is that the number of core questions needs to be decreased by four times as much to accommodate the same number of new questions.

For example, to accommodate sixteen new questions in the one-thirds design requires turning eight questions previously asked of all persons into theme questions. In the two-thirds design it requires turning thirty-two previous questions into theme questions. See the diagram for an example of this. If it is important to retain a large number of core questions it may be worth exploring intermediate designs, in which some themes are assigned to two forms and some to only one.

2.1 Two illustrative designs using three thematic forms


One-third thematic form design
Theme A (8 items from last Census)
Core items
form 1
Theme B (8 new items)
Core items
form 2
Theme C (8 new items)
Core items
form 3
Two-thirds thematic form design
Theme A (16 items from last Census)
Theme B (16 items from last Census)
Core items
form 1
Theme A (16 items from last Census)
Theme C (16 new items)
Core items
form 2
Theme B (16 items from last Census)
Theme C (16 new items)
Core items
form 3

3. ESTIMATION IN A THEMATIC FORM CENSUS

3.1 The long-form / short-form paradigm

The items for a given theme can be viewed as a sample survey of 'long forms'. The full set of Census returns is the survey frame, and the core items are the 'short form' data collected from the whole Census. This core item data provides a very rich set of auxiliary data to be used in estimation from the long-form survey.

3.2 Estimating for a single theme – the weighting approach

The weighting approach attaches a weight to every dwelling that responded for the theme. Estimates for any table involving theme items are obtained by weighted aggregation of the unit data. The simplest approach would be to give each respondent their 'selection weight', the inverse of the probability of a dwelling getting that particular form. In the one-third design, for example, the selection weight of all units is 3.

This simple approach can be improved upon by calibrating the weights to reproduce the actual Census counts for various core items. This can be done at quite a fine geographic level, as was done in estimating for the long-form / short-form Census in Statistics Canada (2002). The standard approach in the ABS for weighting in the presence of rich benchmark information is to use generalised regression (GREG) estimation. Given that both dwelling and person level tables are important, the weighting can be applied at dwelling level, with the weights calibrated to both person and dwelling benchmarks.

3.3 Accuracy of small area estimates

A key issue with moving to a thematic form approach is whether the available sample will give sufficiently accurate estimates for the finely classified items and small geographic areas that may be of interest to users of the data.

For counts of people in a category, an idea of the standard errors available from a sample of the data can be obtained under the assumption that individuals are selected to receive the theme questions independently and with equal probability π. A category with a true count of n will be estimated by $\hat{n} = m/\pi$, for m the observed count in the category for the dwellings responding to the theme.

This set-up leads to the standard error approximation $\mathrm{SE}(\hat{n}) \div \sqrt{n\frac{(1-\pi)}{\pi}}$. If $\pi = \frac{1}{3}$ this gives $\mathrm{SE}(\hat{n}) \div \sqrt{2n}$, while for $\pi = \frac{2}{3}$, $\mathrm{SE}(\hat{n}) \div \sqrt{n/2}$. Thus the two-thirds design gives half the standard error of the one-thirds design. Indicative standard errors and relative standard error (RSE%) arising for estimates of categories of various sizes are given in table 3.1.

3.1 Standard errors and RSE% for counts under two designs

	Size of true count in category				
Thematic form design	*n=8*	*n=50*	*n=200*	*n=800*	*n=3,200*
one-third	4 (50%)	10 (20%)	20 (10%)	40 (5%)	80 (2.5%)
two-thirds	2 (25%)	5 (10%)	10 (5%)	20 (2.5%)	40 (1.3%)

Another likely use of Census data is to obtain a count for a subcategory as a proportion of the category count. Suppose that of the n_c people in a category c we have n_{ct} in a subcategory t, and the corresponding counts for the theme data are m_c and m_{ct}. The estimate of $p_{ct} = \frac{n_{ct}}{n_c}$ is then $\hat{p}_{ct} = \frac{m_{ct}}{m_c}$, with an estimate of standard error given by

$$\mathrm{SE}\,(\hat{p}_{ct}) \div \sqrt{\frac{p_{ct}(1-p_{ct})}{n_c}\,\frac{(1-\pi)}{\pi}} \qquad (1)$$

(See Appendix A for proof.)

Indicative standard errors under the two designs for various proportions and category counts are shown in table 3.2. It will be important to discuss with users the extent to which standard errors of these magnitudes will be fit for the various applications they make of Census data.

3.2 Standard errors (in percentage points) for proportions under two designs

		Size of count in core category				
Thematic form design	*Proportion in theme category*	*n=50*	*n=200*	*n=800*	*n=5,000*	*n=20,000*
one-third	2%	2.8	1.4	0.6	0.3	0.1
	5%	4.4	2.2	1.0	0.4	0.2
	10%	6.0	3.0	1.3	0.6	0.3
	20%	8.0	4.0	1.8	0.8	0.4
	50%	10.0	5.0	2.2	1.0	0.5
two-thirds	2%	1.4	0.7	0.3	0.1	0.1
	5%	2.2	1.1	0.5	0.2	0.1
	10%	3.0	1.5	0.7	0.3	0.2
	20%	4.0	2.0	0.9	0.4	0.2
	50%	5.0	2.5	1.1	0.5	0.3

The estimates $\hat{n}$ and $\hat{p}_{ct}$ above are unweighted, and the standard error measures do not take account of any clustering in the sample due to all persons in a dwelling receiving the same form type. The tabulated standard errors are thus indicative only; if the weighted estimates are produced a better estimate of standard error can be obtained by applying replication methods or a linearisation approach (Bell, 2000).

3.4 Issues with a weighting approach to estimation

Estimates from a weighting approach can only approximately achieve the quality indicated by the above standard errors. In practice there may be differential non-response to the theme questions for different types of respondents, leading to bias. The weighting approach also gives estimates for tables by core items based on units reporting the theme data, rather than the accurate counts available from the Census itself. The weighting process will only be able to control this for a small number of core variables.

Under the weighting approach, estimates are obtained as weighted aggregates from the subset of dwellings that reported a theme. Estimates of core item totals are not guaranteed to agree with the counts obtained from the whole Census. Agreement can be achieved for a set of benchmark counts by calibrating the weights of the theme data so that they add to the Census counts. In fact this is the key reason for calibrating weights – any standard error benefits for estimates of proportions with a given theme response are likely to be marginal.

Unfortunately there are limits to how many benchmark counts can be used, and there will remain many potential tabulations of core items for which estimates based on the theme data will disagree with the Census counts.

A related issue is the need for weights to be used specifically for tables involving a given theme. In the situation where there are three themes each would need its own set of weights. In the two-thirds design some tables involving items from different themes can only be estimated by using that third of the dwellings for which those two themes were reported together. If such cross-theme tables are required they would effectively require a further weight.

There is a concern that both multiple weights and the potential for disagreement between estimates of core items may be confusing to users. The confusion might be lessened if the themes were presented for output purposes as separate large surveys. For instance, in the context of a survey an exact match to all population totals is not expected, and weights and standard errors are accepted as normal.

4. IMPUTATION IN A THEMATIC FORM CENSUS

4.1 Imputation as an estimation method

The issues of consistency and potential user confusion discussed above for the weighting approach provide the motivation to seek an alternative approach to producing estimates from a thematic Census. This section explores using imputation as an estimation method.

The idea of this approach is to treat any themes not collected for a dwelling as missing data, and to impute them on the basis of the information that was collected. Once all this missing data is filled in, estimates for any category can be obtained by unweighted aggregation, treating the imputed values as actual responses.

This approach would overcome the problem of inconsistency between tables from core data and different themes. Tables would all be consistent and the tabulation of theme items would be just as straightforward as for any other item. The only complexity in output would be the need to provide accuracy measures for the resulting estimates.

4.2 A simple hot-deck imputation approach

The ‘hot-deck’ approach fills in missing data for a ‘recipient’ unit from a ‘donor’ unit for which the data was reported. In the simple hot-deck the donor is chosen as a close match to the recipient, as measured using items that both units reported.

One approach to defining a close match is to define ‘imputation classes’ – these are categories defined using the core items within which units are considered to be similar enough to allow imputation. If there are multiple donors in a recipient’s imputation class, the donor can be chosen at random, or as the closest geographically (or in some other sense). It is usual to set a limit on how many times an individual donor is used.

If there are no donors in a recipient’s imputation class, a larger imputation class can be substituted. Imputation based on a predefined hierarchy of imputation classes is known as ‘hierarchical hot-deck’ imputation (David, Little, Samuhel and Triest, 1986).

The data to be imputed needs to make sense with the reported data of the recipient. This leads to only choosing a donor for which the resulting imputed unit passes a set of edit checks. Setting up edit checks that will ensure the data makes sense is a major task in imputation. Editing may be done after the donor is chosen, with a new donor sought only if the edits are failed. Alternatively, a number of potential donors can be edited before choosing which to impute.

4.3 Imputation in the CANCEIS package

The CANCEIS package (Bankier, Poirer and Lachance, 2001) was created by Statistics Canada to perform editing in their Census. The package makes a number of interesting decisions in implementing hot-deck imputation.

First, Statistics Canada impute data for a whole dwelling, rather than imputing person by person. This restricts how detailed the definition of imputation classes can be, since they have to match the dwelling shape (especially the number of occupants), and so there may not be much scope to include a variety of person level items for each person in the dwelling.

As a result, they use large imputation classes. Within the imputation class, a sizeable group of potential donors are chosen based on closeness in a geographic sort order. These are then whittled down according to how much changing of data would be required to make the resulting unit pass edits, and also using a 'distance' function. The distance is calculated as a weighted total of the differences between corresponding items on the recipient and the potential donor. Considerable thought may be needed in designing this distance function – choosing which items to include and the weights to give to a difference in an item.

The result is a handful of potential donors for the recipient, along with a distance for each which can be used in guiding which to choose. A sensible choice is the 'nearest neighbour', but a random choice among the potential donors might be used. The number of times each donor has been used already may also be taken into account.

4.4 Properties of estimates under simple imputation approach

The imputation approaches described above were developed for imputing data that is missing due to non-response. In that setting, it is not clear what distribution the imputed values should take; for example, it is not necessarily a problem if the imputed data does not have the same average behaviour as non-imputed data.

In the thematic form situation the missing data is missed by design. The objective of imputation is to produce estimates – and the resulting figures can be compared to valid estimates from a weighting approach. The simple imputation approach has potential to give estimates for some categories that are markedly inferior to those available from the weighting method.

Estimates from simple imputation are in fact 'synthetic'. For example, the imputed data within a collector's district (CD) may come from similar units from outside that CD. So if that CD had specific characteristics with regard to a theme item, those may be less apparent in the imputed data, being diluted by the characteristics of nearby CDs.

This happens for non-geographic items as well. For example, suppose that the imputation classes and distance function take little account of whether a person is born in Greece. A user tabulating a theme item for Greek-born people in Melbourne will be given estimates that include imputed data from ‘similar’ people who are almost all not Greek. Any peculiarity of the Greek community will only appear in the non-imputed portion of the dataset, and the figures will be diluted (i.e. biased) by the imputed data.

With the large sample size available, allowing estimates at fairly high levels of aggregation to be ‘synthetic’ will give inferior estimates to those from the weighting approach. On the positive side, the imputation approach gives consistent estimates for all categories; in this it is much more successful than the weighting approach which is limited to achieving consistency with a limited number of benchmark category counts. The next section explores an imputation approach that aims to provide estimates for theme items that are consistent with those available from the weighting approach.

5. A BALANCED IMPUTATION APPROACH

5.1 Imputing to preserve estimates

The imputation methods discussed above are not limited to providing a single impute for each unit. In fact, they quite naturally provide multiple potential imputes, all passing the necessary edit checks, and each with a distance measure to inform which impute should be preferred.

The concept of balanced imputation is to choose from among the set of imputes available in such a way that the imputed data as a whole fulfills certain criteria. Specifically, the chosen imputes should give counts that are appropriately close to estimates based on the theme data only. How closely a given estimate should be reproduced by the imputed data will be informed by a simple estimate of the estimate's standard error.

The method allows for balancing of the imputes to be performed with reference to a large set of tables. The chosen set needs to include a wide variety of different tabulations, as accuracy needs to be acceptable for the many kinds of tables that could be required by Census users. Theme items classified very finely (e.g. occupation) may contribute via tables at only the broadest geographic levels. A broad classification based on the same item may be used cross-classified by a number of other variables. Determining an appropriate set of tables to use in balancing could be a major undertaking; rather than focus on this, the approach proposed is to balance with respect to a large ensemble of tables, using criteria that will be relatively unaffected by cells that have high sampling error or where the range of potential imputes all reproduce the estimates equally well.

Balanced imputation resembles in some ways the balanced sampling methods of Deville and Tillé (2004), in that the aim is to make selections (in our case, from among the potential imputes) in a way that achieves overall target totals. There are major differences, however; particularly the large number of constraints in the imputation setting (including that we must select exactly one impute for each unit) and the fact that the target totals do not need to be reproduced exactly.

5.2 A theoretical setting for balanced imputation

Distribution of imputed values before and after choice of imputes

We can think of every unit $i \in \mathrm{U}$ (the population) as having its theme item values (vector $\boldsymbol{t}_i$) drawn from a distribution T_i with density $\mathrm{p}^{\mathrm{T}}(\boldsymbol{t}_i \,|\boldsymbol{x}_i)$ conditional on the unit's core item values (vector $\boldsymbol{x}_i$). For each unit $i \in \mathrm{U}^{\mathrm{T}}$, the set of units which reported this theme, we have observed the theme values; this provides a great deal of information about relationships between core and theme values. The unit used could be either dwelling or person; person is used for the evaluation presented in this paper.

Imputation is an attempt to obtain a draw from the conditional distribution for each unit $i \in \mathrm{U}^{\mathrm{I}}$ (the set of n^{I} units not reporting this theme). Hot-deck imputation provides multiple draws $\boldsymbol{t}_i^k$ for $k = 1, \ldots, K_i$ from a related conditional distribution H_i (with conditional density $\mathrm{p}^{\mathrm{H}}(\boldsymbol{t}_i \,|\, \boldsymbol{x}_i)$) that was observed empirically among the theme data units with similar $\boldsymbol{x}_i$ values in some sense. The final impute is chosen from these potential imputes. The simple hot-deck would choose one at random, or based on a distance function. Balanced imputation chooses the impute so as to improve the outcomes of imputation at an aggregate level, the hope being that this will also modify the effective distribution being used for imputation of the individual unit (from H_i and hopefully towards T_i).

Predicting the outcome of imputes that have not yet been considered

The key information required for choosing among the potential imputes is a prediction, for each table cell in an ensemble of tables, of what the cell count after imputation would be *if the choice of imputes was made without reference to its effect on this table cell*. Potential imputes are judged by whether they move this value closer to or further from the target cell value.

For the purpose of predicting the outcome of such an *unbalanced* imputation for a unit i, it is assumed that potential impute k will be chosen with a probability ϕ_{ik}. This set of imputation probabilities $\boldsymbol{\phi} = \{\phi_{ik}\}_{i \in \mathrm{U}^{\mathrm{I}}, k=1,\ldots,K_i}$ is used to predict the overall distribution of imputed cell counts under unbalanced imputation of the units not yet imputed. Choosing an impute replaces the imputation probabilities for the unit i to describe the actual outcome, so that after imputation exactly one of $\phi_{i1}, \ldots \phi_{iK_i}$ is one and the others are zero.

In the algorithm developed here, the imputation choice for many units will be determined by aggregate measures (such as the predicted fit of imputed counts to estimates). For other units the aggregate measures will not strongly prefer a particular impute, in which case an impute will be chosen from donors with a better match on core values, or donors that have been used fewer times in imputing other units. It may be adequate to reflect the uncertainty in these choices by setting $\phi_{ik} = K_i^{-1}$; that is, each potential impute is judged equally likely to be chosen.

5.3 Defining the ensemble of Census tables requiring balancing

We wish to choose imputes so as to obtain good fit for a specified ensemble Q of Census tables. Clearly this set of tables needs to be limited to avoid extremely finely classified tables. Each table in the ensemble will cross-classify a small number of Census items, but the items themselves would not all be used at the finest levels of classification (i.e. not fine geography by fine age). The levels of classification available for each item will be numbered from 0 upwards, with 0 signifying the table is not classified by this item, 1 being the broad category (often only two or three values) and so on. Cells which appear in more than one table are listed only once, to minimise redundant calculation and to avoid an inappropriate emphasis on those cells in balancing.

Note that the classifications do not have to cover all possible values: for example, the finer classifications may omit any categories that are the same as a category in a broader classification, so as not to compute the same figures in multiple tables.

Geography will play a major role in defining tables in the ensemble to be used. The level of geography will be denoted by $l = 0, \ldots, L$ with level 0 being the whole imputation region within which imputation occurs (perhaps 200,000 units). The algorithm ensures that at any point a single geographic area at each geographic level is required for the calculations. This avoids retaining totals for all tables in RAM memory. A side effect of this is that imputation at dwelling level will only consider the effect of imputation on a single geographic classification that the dwelling contributes to (although the choice made will be accounted for in imputing other dwellings).

Consider each table in the ensemble Q as being defined by the level of geography l and a table number $q = 0, \ldots, Q_l - 1$. The corresponding table descriptions Γ_q consist of a string of small integers, one for each Census item other than geography, with each integer giving the level of classification at which table q uses that item. Note that these descriptions are the same for the same numbered table at each geographic level, but the broader geographic levels will have more tables. Thus $Q_0 \geq Q_1 \geq \ldots \geq Q_L$, with the lowest numbered tables being the broader ones (since they can afford to be classified by fine geography).

The number of tables at imputation region level (Q_0) is expected to be quite large. Tables at this level could include totals of all theme items (at their finest classification) across the whole imputation region, totals of many theme items classified by individual core items at their finest level, and cross-classifications of multiple theme items and core items.

5.4 Measuring the impact of potential imputes on aggregates

Consider table q at geographic level l with cells indexed by a geographic area g (numbered from 0 to $G_l - 1$), a core cell c (numbered from 0 to $C_q - 1$) and a theme cell t (numbered from 0 to $T_q - 1$). For notation, categories will be indexed by the variables *lqgct*. The total number of cells in all tables will be denoted n^{CELLS}.

The number of persons n_{gc}^{lq} in marginal category *lqgc* is known as this category uses only core items. The table obtained by cross-classifying by theme value t is only available for the theme data. Let m_{gc}^{lq} be the count of theme units in marginal category *gc* and let m_{gct}^{lq} be the number taking theme value t.

A ratio estimate of the number of units in category *lqgct* is

$$\hat{n}_{gct}^{lq} = n_{gc}^{lq} \frac{m_{gct}^{lq}}{m_{gc}^{lq}} \tag{2}$$

Using (1), the standard error of $\hat{n}_{gct}^{lq}$, for $p_{gct}^{lq} = n_{gct}^{lq} / n_{gc}^{lq}$, is approximately

$$\mathrm{SE}(\hat{n}_{gct}^{lq}) \doteq \sqrt{n_{gc}^{lq} p_{gct}^{lq} (1 - p_{gct}^{lq}) \frac{(1-\pi)}{\pi}} \tag{3}$$

This is estimated using

$$\hat{s}_{gct}^{lq} = \sqrt{n_{gc}^{lq} \left(\frac{m_{gct}^{lq} + T_q^{-1}}{m_{gc}^{lq} + 1} \right) \left(1 - \frac{m_{gct}^{lq} + T_q^{-1}}{m_{gc}^{lq} + 1} \right) \frac{(1-\pi)}{\pi}} \tag{4}$$

with the T_q^{-1} adjustment applied to avoid extreme sample proportions.

The value $\hat{n}_{gct}^{lq}$ gives a target total for the category *lqgct* after imputation – but this target will not be able to be met exactly, given the competing priority of other tables and categories. The value $\hat{s}_{gct}^{lq}$ is used as a scaling factor against which the importance of a difference between the imputed count $\ddot{n}_{gct}^{lq}$ and the estimate $\hat{n}_{gct}^{lq}$ can be measured.

This allows a normalised measure of the fit of the expected imputes based on a given set of imputation probabilities $\boldsymbol{\phi}$. Write J_{ikt}^{lq} as the contribution (0 or 1) of unit i to theme category t of table q at geographic level l when using the kth potential impute for unit i's theme values. The predicted contribution of unit $i \in \mathrm{U}^{\mathrm{I}}$ to this category (or the actual contribution if i has already been imputed) is then

$$\bar{\mathrm{J}}_{it}^{lq}(\boldsymbol{\phi}) = \textstyle\sum_k \phi_{ik} \, \mathrm{J}_{ikt}^{lq} \tag{5}$$

The predicted total for category *gct* after imputing using probabilities ϕ is given by

$$\ddot{n}^{lq}_{gct}(\phi) = m^{lq}_{gct} + \sum\nolimits_{i \in \mathrm{U}^{\mathrm{I}}:\ i \text{ in category } lqgc} \bar{\mathrm{J}}^{lq}_{it}(\phi) \tag{6}$$

The normalised fit of this total to the estimated value $\hat{n}_{gct}$ is then given by

$$\mathrm{F}^{lq}_{gct}(\phi) = \frac{\ddot{n}^{lq}_{gct}(\phi) - \hat{n}^{lq}_{gct}}{\hat{s}^{lq}_{gct}} \tag{7}$$

This fit measure can only be affected by units which have a probability above zero and less than one of being imputed to the relevant table cell i.e. $0 < \bar{\mathrm{J}}^{lq}_{it}(\phi) < 1$. Writing I() for the indicator function (taking value 1 if the expression in the brackets is true and 0 otherwise), the maximum and minimum fit values that can be achieved are given by

$$\mathrm{F}^{\mathrm{HI}\,lq}_{gct}(\phi) = \sum\nolimits_{i \in \mathrm{U}^{\mathrm{I}}:\ i \text{ in category } lqgc} \mathrm{I}\left(\bar{\mathrm{J}}^{lq}_{it}(\phi) > 0\right)$$

and
$$\mathrm{F}^{\mathrm{LO}\,lq}_{gct}(\phi) = \sum\nolimits_{i \in \mathrm{U}^{\mathrm{I}}:\ i \text{ in category } lqgc} \left(1 - \mathrm{I}\left(\bar{\mathrm{J}}^{lq}_{it}(\phi) < 1\right)\right) \tag{8}$$

The impact on category *lqgct* of any specified choice of imputes can be gauged in terms of the change to the normalised fit. This paper will go on to outline an algorithm that imputes units one at a time so as to improve the normalised fit over all the categories of an ensemble of tables of Census estimates.

5.5 A balanced hot-deck imputation procedure

Objective of balancing algorithm

The aim of the balancing algorithm is to provide a set of imputes for which the fit measures for all cells of all tables in *Q* are small. This suggests treating the maximum absolute value of the fit measure $\mathrm{F}^{lq}_{gct}(\phi)$ across all *lqgct* as the objective function to be minimised. This paper opts to use the alternative objective of minimising the following mean-squared fit measure, where the mean is taken across all cells in the ensemble of tables *Q*:

$$\mathrm{MSF}(\phi) = \frac{1}{n^{\mathrm{CELLS}}} \sum\nolimits_{lqgct} F^{lq}_{gct}(\phi)^2 \tag{9}$$

This measure of fit emphasises cells with poor fit, while being easy to compute. The description below is based around choosing the impute at any stage that most reduces this mean squared fit value (though a couple of variations on this are suggested). Note that the algorithm provides a heuristic method for identifying the best impute to choose for each unit in turn. It does not guarantee the lowest possible value for either of these overall fit measures.

Overview of algorithm

The algorithm considers each unit requiring imputes in turn, after randomly ordering the units. For each potential impute it evaluates the effect of using that impute and compares it to the effect of using each other impute. As the criterion for this evaluation it uses the change in the overall mean-squared fit measure. The effect is evaluated of each potential impute on each cell of the whole suite Q of tables; equation (9) is then used to summarise this effect across all tables. The impute that minimises the mean-squared fit is chosen as the final impute – although if there are multiple potential imputes with similar effects an appropriate random choice can be made from among these acceptable imputes.

Step 1: Evaluate balance of predicted imputes

Assign $\boldsymbol{\phi}$ the initial imputation probabilities, ensuring that $\Sigma_k \, \phi_{ik} = 1$ for all i.

Sort the whole population U hierarchically by geography, and on a single pass through the dataset, calculate:

- the aggregates: n_{gc}^{lq} , m_{gc}^{lq}, m_{gct}^{lq} and $\ddot{n}_{gct}^{lq}(\boldsymbol{\phi})$, and
- the measures: $\hat{n}_{gct}^{lq}$, $\hat{s}_{gct}^{lq}$, $\mathrm{F}_{gct}^{lq}(\boldsymbol{\phi})$, $\mathrm{F}_{gct}^{\mathrm{LO}lq}(\boldsymbol{\phi})$and $\mathrm{F}_{gct}^{\mathrm{HI}lq}(\boldsymbol{\phi})$

for all cells gct in each table $lq \in Q$. The values are stored in a separate dataset for each geographic level, to be accessed as needed in later steps.

Step 2: Assign random number and order of imputation

For later steps only the imputation units $i \in \mathrm{U}^{\mathrm{I}}$ are used. These units are partitioned into a^{PH} portions of approximately equal size, numbered $\psi = 1, \ldots, a^{\mathrm{PH}}$. The portions are treated one at a time, with all units in each portion being processed in geographic order.

This partition allows computation to proceed for one geographic classification at a time, while ensuring that each geographic region has some units imputed early and some late in the procedure. If the final implementation is at dwelling level there is another benefit, since there will be dwellings in one imputation region contributing to other regions. In this situation, imputing a portion from each imputation region in turn will ensure that the effect of imputing the multi-region dwellings in one region can be taken into account in imputing each other region.

The units are each assigned a random number r_i on $[0,1)$ (denoting the range $0 \le r_i < 1$). Portion ψ is assigned units i with random numbers in $[(\psi - 1)/a^{\mathrm{PH}}, \psi/a^{\mathrm{PH}})$. For most units the random number is drawn from a Uniform$(0,1)$ distribution. Certain potentially difficult units (such as large or multi-geography dwellings in dwelling-level imputation) may have their random numbers drawn from a smaller range (e.g. $\mathrm{Uniform}(0, a^{\mathrm{PD}}/a^{\mathrm{PH}})$) to ensure that they do not get assigned to the later portions.

Separate datasets are created for each portion, with units in each dataset sorted hierarchically by geography and within each geographic area by random number r_i. This is the order in which the units will be imputed: renumber the units so that they take successive identifiers $i = 1, \ldots, n^{\mathrm{I}}$ in this order.

Proceed to do step 3 for each portion ψ.

Step 3: Impute for each unit in order

For each unit i in turn within portion ψ do steps 3a to 3f.

Step 3a: Read in new geographic areas

Do the following for each geographic level l for which i is the first unit (in this portion) in a geographic area g:

For each table $q = 0, \ldots, Q_l - 1$, read in (from the appropriate dataset)

- the aggregates: n^{lq}_{gc} , m^{lq}_{gc}, m^{lq}_{gct} and $\ddot{n}^{lq}_{gct}(\phi)$, and
- the measures: $\hat{n}^{lq}_{gct}$, $\hat{s}^{lq}_{gct}$, $\mathrm{F}^{lq}_{gct}(\phi)$, $\mathrm{F}^{\mathrm{LO}lq}_{gct}(\phi)$and $\mathrm{F}^{\mathrm{HI}lq}_{gct}(\phi)$

for $c = 0, \ldots, C_q - 1$ and $t = 0, \ldots, T_q - 1$.

Adjust these measures for any contribution from multi-geography units that were imputed since this new geographic area was last considered. These units will have been imputed based upon their contribution to another geographic area, but the aggregates and measures for this area need to be adjusted to account for the choices made. (Details are omitted as this does not arise in the person-level imputation evaluated in this study, although it could if both place-of-enumeration and place-of-usual residence tables were included in Q.)

Step 3b: Calculate fit change measures for individual tables

For $l = 0, \ldots, L$ and $q = 0, \ldots, Q_l - 1$

Define $g(l, i)$ as the level l geographic area of unit i

Define $c(q, i)$ as the core category unit i contributes to in table q

For all theme categories $t = 1, \ldots, T_q$ of table q

Compute for each potential impute $k = 1, \ldots, K_i$ J^{lq}_{ikt}, the contribution to table q theme category t of potential impute k of unit i

hence evaluate $\bar{\mathrm{J}}^{lq}_{it}(\phi)$ using (5).

Now for $t = 1, \ldots, T_q$ and $k = 1, \ldots, K_i$ calculate the change in fit:

$$f^{lq}_{ikt}(\phi) = \frac{\mathrm{J}^{lq}_{ikt} - \bar{\mathrm{J}}^{lq}_{it}(\phi)}{\hat{s}^{lq}_{g(l,i)c(q,i)t}} \tag{10}$$

The change in squared fit depends upon the current fit of cell $g(l,i)c(q,i)t$ of table q at level l. The change in squared fit for this table and cell is given by

$$\begin{aligned} \delta^{lq}_{ikt}(\phi) &= \left(\mathrm{F}^{lq}_{g(l,i)c(q,i)t}(\phi) + f^{lq}_{ikt}(\phi)\right)^2 - \mathrm{F}^{lq}_{g(l,i)c(q,i)t}(\phi)^2 \\ &= \left(2\mathrm{F}^{lq}_{g(l,i)c(q,i)t}(\phi) + f^{lq}_{ikt}(\phi)\right) f^{lq}_{ikt}(\phi) \end{aligned} \tag{11}$$

Step 3c: Calculate overall measures of fit for each impute

The change in the total squared fit from choosing potential impute k is then

$$\Delta_{ik}(\phi) = \frac{1}{n^{\mathrm{CELLS}}} \sum_l \sum_q \sum_t \delta^{lq}_{ikt}(\phi) \tag{12}$$

Alternative measures could be proposed here. It would be possible to only sum changes to fit measures that were outside a target range of values $(-\alpha^{\mathrm{FT}}, \alpha^{\mathrm{FT}})$. This would give the alternative measure of fit change:

$$\Delta^{\mathrm{FT}}_{ik}(\phi) = \frac{1}{n^{\mathrm{CELLS}}} \sum_l \sum_q \sum_t \delta^{lq}_{ikt}(\phi)\, \mathrm{I}\left(\left|\mathrm{F}^{lq}_{g(l,i)c(q,i)t}(\phi)\right| > \alpha^{\mathrm{FT}}\right) \tag{13}$$

A variation on this approach would be to weight the changes according to the change required to bring the fit to the target value, measured as a proportion of the total possible change in that direction. This would give the weights

$$\omega^{lq}_{it}(\phi) = \begin{cases} \dfrac{-\alpha^{\mathrm{FT}} - \mathrm{F}^{lq}_{g(l,i)c(q,i)t}(\phi)}{\mathrm{F}^{\mathrm{HI}\,lq}_{g(l,i)c(q,i)t}(\phi) - \mathrm{F}^{lq}_{g(l,i)c(q,i)t}(\phi)} & \text{if } \mathrm{F}^{lq}_{g(l,i)c(q,i)t}(\phi) < -\alpha^{\mathrm{FT}} \\[2ex] \dfrac{\mathrm{F}^{lq}_{g(l,i)c(q,i)t}(\phi) - \alpha^{\mathrm{FT}}}{\mathrm{F}^{lq}_{g(l,i)c(q,i)t}(\phi) - \mathrm{F}^{\mathrm{LO}\,lq}_{g(l,i)c(q,i)t}(\phi)} & \text{if } \mathrm{F}^{lq}_{g(l,i)c(q,i)t}(\phi) > \alpha^{\mathrm{FT}} \\[2ex] 0 & \text{otherwise} \end{cases}$$

with the resulting weighted measure of fit change

$$\Delta^{\mathrm{FW}}_{ik}(\phi) = \frac{1}{n^{\mathrm{CELLS}}} \sum_l \sum_q \sum_t \omega^{lq}_{it}(\phi)\, \delta^{lq}_{ikt}(\phi) \tag{14}$$

Some experimentation may be required to determine how well each of these fit measures performs, in terms of the overall fit of the eventual imputed dataset. The description below assumes that $\Delta_{ik}(\phi)$ is being used.

Step 3d: Choose the impute

In practice, a number of the potential imputes may be almost equally good as measured by their effect on the total squared fit. Define the set of acceptable imputes A_i as those sufficiently near to the best impute available for unit i, as defined by a parameter α^{AF}:

$$A_i = \left\{ k : \left(\Delta_{ik}(\phi) - \min_{k^*} \left(\Delta_{ik^*}(\phi) \right) \right) \le \alpha^{\mathrm{AF}} \right\} \tag{15}$$

Choose at random with probability proportional to ϕ_{ik} from among the acceptable imputes whose donors have been used the least number of times.

At this stage, if this was a multi-geography unit it needs to be stored so that the impact of the impute can be accounted for when imputation is next considered for the other geographic areas the unit contributed to (details omitted).

Step 3e: Adjust aggregates and measures

Modify the set of aggregates and measures to account for the imputation of unit i.

Also, there may be parameters relating to the level of fit and fit change that are to be aimed for in imputation. These can be updated at this point.

Step 3f: Output aggregates and measures if the last unit in a geographic area

These are stored as they will be needed when the geographic area is revisited in the next portion of units.

Step 4: Output summary measures of fit and adjust parameters

Diagnostic information about fit can be produced before the imputation and again at the end of imputing each portion. The following information would be of interest.

Distribution of fit measures $\mathrm{F}_{gct}^{lq}(\phi)$. These could take the form of a SAS PROC UNIVARIATE i.e. overall mean and spread measures, key percentiles and extreme values. These could be classified by geographic level l.

Summary of distribution of fit measures across tables involving each theme item.

Number of table cells with $\mathrm{F}_{gct}^{lq}(\phi) < -\alpha^{\mathrm{FT}}$ and total of fit measures for these cells, and similarly for cells with $\mathrm{F}_{gct}^{lq}(\phi) > \alpha^{\mathrm{FT}}$.

For each geographic level, plots of fit measures (perhaps ignoring those within a range of zero to avoid too dense a plot). These could be spread out across the horizontal axis by estimate size or grouped by table number.

Adjusting parameters used in the algorithm

It may be sensible to start with conservative values of the various parameters defining the algorithm. The evaluation at the end of each portion provides an opportunity to update the parameters to be used in imputing the next portion for that imputation region. Thus the target fit value a^{FT}could be lowered if it became apparent that the original target will be easily achieved.

Note on the role of randomness in balanced imputation

Balanced imputation will often make a purposive choice from among the potential imputes, but the impute chosen will depend on which other units have already been considered. Even with such a purposive choice, randomness remains in the procedure since the units are imputed in a randomised order.

For example, suppose that the predicted counts show a deficit of persons in a theme category “English ancestry” within the core category “born in England”. Balanced imputation will tend to correct this by imputing into this category when the opportunity arises, until the deficit becomes small enough. This deficit will therefore have a greater influence on the imputes chosen for units considered earlier in the process. Nevertheless, the set of units imputed to “English ancestry” in this way is effectively randomised via the random sort order in which the units are imputed.

5.6 Properties of estimates under a balanced imputation approach

The balanced imputation approach described above will, if successful, provide a set of imputed Census counts that are very compatible with the best estimates available for the ensemble of tables Q used in balancing. The imputation will also have the appealing properties of consistency across tables that were discussed in Section 4.

An important question is: for what types of tables *not* included in Q will the balanced imputation also provide good estimates? Conversely, in what situations and for what sort of tables will the balanced imputation be markedly inferior to an estimation approach targeted at the same table?

It is clear that estimates not covered in Q are still synthetic, but they should be improved synthetic estimates compared to a simple hot-deck imputation. For example, if Greek-born people turn out to have higher values for a theme item for a broad geographic area, this will be represented in the data by higher imputed values for Greek-born people across that area, and hence in the component CDs. Specifically, the imputes applied to a CD with a large Greek-born population will have high values of the theme item. The resulting synthetic estimate is effectively informed by a richer underlying model than would have applied under simple hot-deck imputation; hopefully there will be fewer situations therefore where the imputed value is markedly inappropriate.

The imputation may also be misleading if there are strong relationships between theme and core items that are not represented in the original hot-deck and that do not appear in the set of tables Q. The method presented attempts to allow the ensemble of balancing tables Q to be as large as possible, so as to allow for the possibility of such relationships appearing.

6. COMPARISON OF METHODS USING 2006 DATA

6.1 Assumptions in setting up the evaluation

To enable the project to proceed in a short timeframe, a number of assumptions were made beforehand to limit the scope of the evaluation.

Theme and core items

For the purpose of this study five items from the 2006 Census were designated 'theme items' and 18 were designated as 'core items'. These items were identified in discussion with Census staff. A listing of the data items is presented in table B.1 in Appendix B.

Each dwelling independently had a one-third chance of providing a response to a theme item (the remaining dwellings were assumed to have been given forms containing questions about other themes). An alternative would have been to assume slightly more control over the distribution of theme dwellings by assigning the theme to a third of responding dwellings in each CD, but it is not clear that this level of control is realistic. An alternative option in which the theme items are collected from two-thirds of the dwellings could have also been explored.

Geographic level

In practice imputation could be performed within a state/territory by capital city/balance. For this project the evaluation was limited to a single imputation region, which was all of Hobart. A method that works acceptably for this region is required before exploring its application to the remainder of Australia.

Three lower levels of geography were defined at which estimates were produced. From broadest to finest these were statistical local area (SLA), suburb (SSC) and collector's district (CD). Refer to table 6.1 below for details on the number of categories for each geographic level. These levels are used to demonstrate the properties of estimates, and do not necessarily correspond to the levels at which final Census outputs will be needed. It may be desirable to include an additional level of geography which fits between SLA and the imputation region, for example statistical subdivision (SSD). In this evaluation SSD could not be used because Hobart constitutes a single SSD.

Assignment of donors

The process for providing donors for imputation used a hierarchical hot-deck approach based on defining a hierarchy of imputation classes. Each imputation class consisted of a geographic area cross-classified with a demographic classification 'role' based on sex, age and relationship in dwelling. The levels of hierarchy correspond to the level of geography by which this role is classified.

Within a role, the imputation class was a geographic area for which the number of donor units available exceeded a minimum α^{DM} and a given proportion α^{DC} of the number of recipients. To achieve this, geographic areas appearing successively in the geographic sort order frequently had to be collapsed together into a single imputation class. In doing this, areas within a single higher level geographic unit (e.g. CDs within a Suburb) were collapsed in preference to collapsing across areas from different higher level units.

Each unit requiring imputation was assigned six donor units from which the balanced hot-deck could select a single impute. The values of $\alpha^{\mathrm{DM}} = 10$ and $\alpha^{\mathrm{DC}} = 0.4$ were used. With this constraint on the proportion of donors to recipients, a given donor could not be used for balanced imputation more than (6 / 0.4 =) 15 times. It should be noted the hierarchical hot-deck process used never collapsed across the imputation class, so occasionally the imputation class did not meet the constraints for α^{DM} and α^{DC}. However in all cases the imputation class had at least six donor units.

Note that this imputation may be inferior to a distance-based imputation (perhaps using CANCEIS) in which the suitability of potential donors is based on a more extensive set of variables which do not have to match perfectly between the donor and receiving record. It was infeasible to implement such an approach in the timeframe of this project, but it could be considered for an eventual implementation.

Census data to be treated as giving true values

The true value for each geographic level is assumed known from the Census responding dwellings. For the purpose of this project "Not stated" will be treated as a separate category for core and theme items (other than those filled in by Census processing i.e. sex, age and marital status).

Tables and estimates used in evaluation

The full set of tables used for evaluation will be denoted Q^{E} and the subset used for balancing denoted Q. The tables were classified into 'Detail Level' categories according to the number of variables contributing to the table and the number of table cells. The Detail Level determined the finest level at which the table was involved in the balancing process. A summary of the number of tables and cells is presented in table 6.1 below, while table 6.2 presents a breakdown by Detail Level. Note that Q^{E} contains a few tables at alternative geographic levels or finer classifications than the set Q used in balancing, allowing the performance of the methods to be compared on tables not specifically used to guide the balanced imputation process.

For the tables Q^E, cell estimates based on the one-third theme dwellings were produced for comparison to the true values n_{gct}^{lq}. All imputation was at person level, and geography was on a place of usual residence basis (persons with usual residence outside of Hobart were excluded from the study). The estimates compared are:

- simple estimates obtained by multiplying by 3 the m_{gct}^{lq}, the observed number of theme units taking theme value t;
- ratio estimates $\hat{n}_{gct}^{lq}$, which are used as the target cell values for the balanced imputation;
- average hot-deck, which is the estimate obtained when the units requiring imputation contribute the average cell values of their six assigned imputes. This average is used as the starting point for the balanced hot-deck algorithm;
- balanced hot-deck.

No analysis was performed on a simple hot-deck estimator, obtained when a single donor unit is assigned to each recipient. Clearly the simple hot-deck will be inferior to the average hot-deck described above, which in turn is improved upon by the balanced hot-deck.

6.1 Summary of table cells in Q^E

Geographic Level	*Number of classes*	*Number of tables involved with balancing*	*Number of cells per class*	*Total number of cells for geography*	*Number of cells involved with balancing per class*	*Total number of cells involved with balancing for geography*
Hobart	1	2,360	183,768	183,786	155,967	155,967
SLA	8	1,495	81,930	655,440	56,631	453,048
Suburb	94	441	28,130	2,644,220	9,731	914,714
CD	388	107	2,068	802,384	1,258	488,104

6.2 Classification of table cells by Detail Level

Detail Level	*Number of tables*	*Finest geography used for balancing*
Detailed	123	Not balanced
Fine	865	Hobart
Moderate	1,054	SLA
Broad	334	Suburb
Few cells	107	CD

Ratio estimation of cell totals

The ratio estimate $\hat{n}_{gct}^{lq}$ was produced for each table cell used in the evaluation. For individual cells this estimate may be superior to the cell estimate obtained from a weighted estimator which assigns a single weight to all theme units and applies these weights to obtain a cell estimate. This superiority is most likely to occur for cells at fine geography because the core cell count n_{gc}^{lq}, which is used by $\hat{n}_{gct}^{lq}$ but not the unit-weighted estimator, has highest RSE for fine l. A disadvantage of the estimator $\hat{n}_{gct}^{lq}$ is that it lacks the property of being additive across categories.

Measures of accuracy based on the known Census value

The true error of each estimate was available for this evaluation, since the full Census value was available. Measures of accuracy for the different estimators were obtained by summarising these true errors across a large number of table cells. Sample-based measures of accuracy which account for the household-level clustering were not used in this evaluation. A separate project is required to develop these measures.

Repeated simulations

The above evaluation could be repeated for many different assignments of the population units to 'theme units' and 'core units'. Averaging across the simulations would confirm the properties observed are independent of which units are assigned as 'theme units' and enable thorough evaluation of potential bias for the imputation for specific tables or items. The computing time required to perform the balanced imputation makes such an evaluation impractical. Considering the assessment of the estimators is based on average behaviour across several million table cells, the comparison between estimators obtained in this evaluation should be robust to the assignment of units to 'theme units'.

6.2 Results

Excluding small cells and those involving a "Not stated" or "Not applicable" response

The following analysis of estimator performance excludes cells with $\hat{n}_{gct}^{lq}$ less than or equal to 20 and cells involving a response value of "Not stated" or "Not applicable".

Small cells are excluded because they are less likely to be of interest to the user and they are prone to small standard errors, thereby distorting summary measures expressed relative to the size of SE.

The exclusion of cells featuring a "Not stated" value greatly reduces the spread of the error distribution for the average hot-deck estimator (refer to Appendix C). This result indicates the hot-deck performs comparatively worse for such cells, which is not surprising since the variables which form the imputation classes may have weak relationship with the likelihood of a person providing a response.

It appears likely that "Not stated" values for the core items are a good indicator of "Not stated" values for the theme items. It may be possible to use this fact in designing the imputation classes used in the hot-deck step of the imputation. This could well lead to improvements in fit for these "Not stated" cells.

Considering this, and that cells featuring a "Not stated" or "Not applicable" are unlikely to be of interest to the user, it was decided to exclude such cells from analysis of performance.

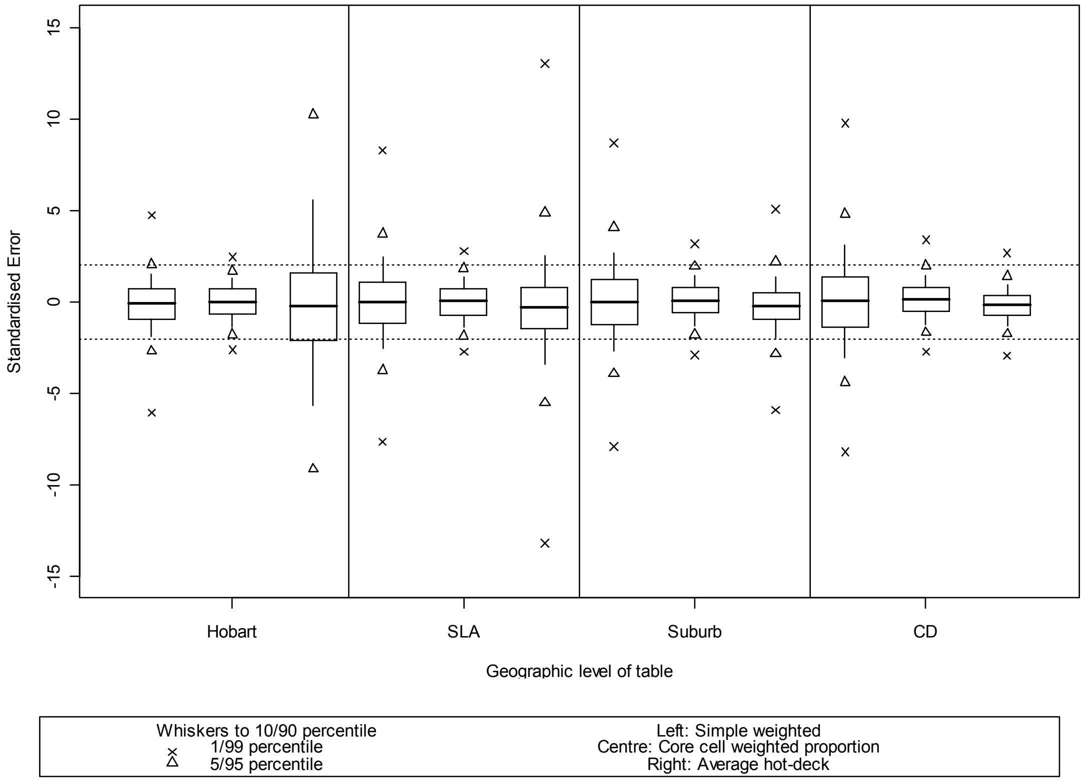
Performance of hot-deck imputation

Graph 6.4 presents boxplots comparing the performance of the average hot-deck imputation with the ratio estimates $\hat{n}_{gct}^{lq}$. The box-plots represent the across-cell distribution of the error from the true cell value n_{gct}^{lq}, relative to the approximate SE of $\hat{n}_{gct}^{lq}$, as given by (4). This relative error will henceforth be referred to as the Standardised Error. The proportion of cells at each geographic level with Standardised Error outside the range (–2,2) is given in table 6.3.

6.3 Proportion of cells with Standardised Error outside range (–2, 2)

Geographic Level	*Simple weighted*	*Ratio estimate* $\hat{n}_{gct}^{lq}$	*Average hot-deck*
Hobart	14.2%	6.1%	47.5%
SLA	27.5%	7.6%	31.4%
Suburb	29.9%	8.0%	15.5%
CD	34.4%	7.8%	5.5%

6.4 Distributions across cells of Standardised Error for three estimators



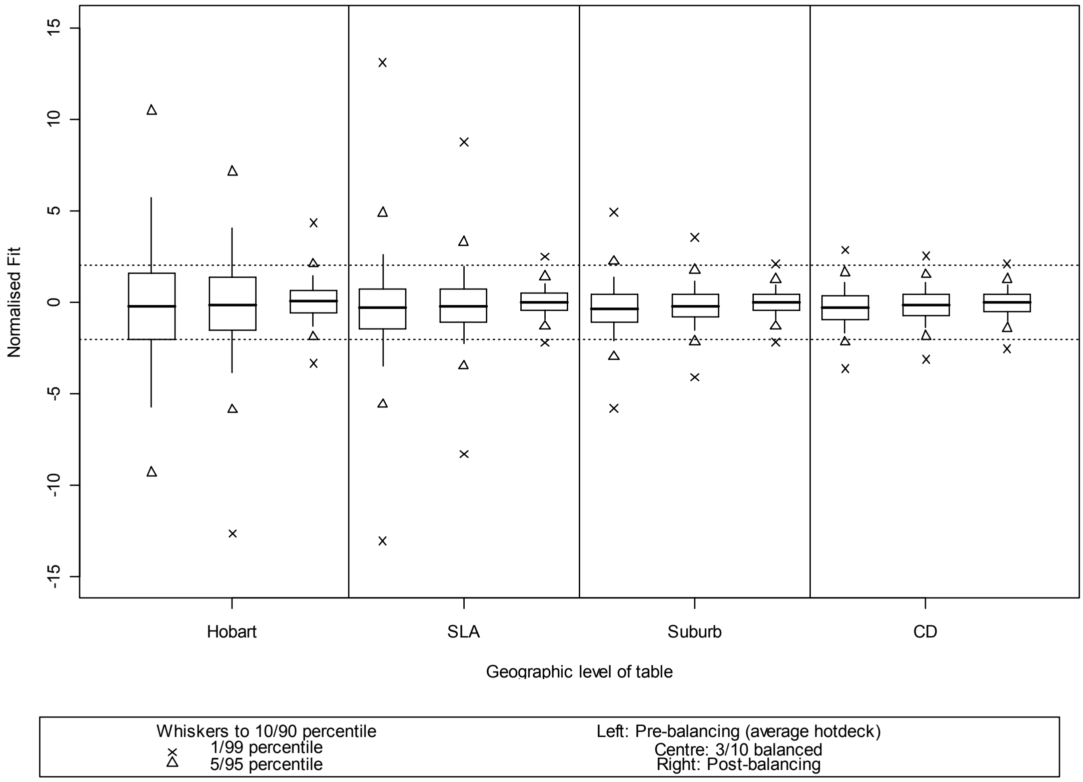
The relative performance of the different estimators varies across the geographic level of the table cell. Compared to the ratio estimates $\hat{n}^{lq}_{gct}$, the simple weighted estimate performs worse for cells at fine geography. This is because the simple weighted estimator does not control for n^{lq}_{gc}, the size of the core cell. The average hot-deck performs worse for cells at the broader geographic levels since any bias inherent in the assignment of donors prior to balancing will accumulate across the finer geographic levels. The Standardised Error will typically increase as this bias accumulates because the Standardised Error measures discrepancy relative to the SE, which is proportional to the square root of the size n^{lq}_{gc} rather than the size of n^{lq}_{gc}.

Gains from balancing process

For the analysis of the performance of the balanced imputation process, the focus is on the difference between the balanced hot-deck value and the ratio estimate $\hat{n}^{lq}_{gct}$. The $\hat{n}^{lq}_{gct}$ provide the target cell value for the balanced hot-deck, so it provides a fairer value for comparison than the actual cell value n^{lq}_{gct}. The difference is measured as a proportion of the approximate SE (4), resulting in the Normalised Fit measure (7).

Graph 6.5 compares the distribution of the Normalised Fit before, during and after balancing, for tables at each geographic level. The middle box-plot in each group shows the distribution after 30% of the recipient units have had a donor assigned through balancing choice. The balancing process is able to get most table cells close to their target values. After balancing, only 3.2% of table cells at CD level have a Normalised Fit outside the range (–2,2). For tables relating to all of Hobart, the proportion is reduced from 47.7% prior to balancing, to 9.9% post-balancing.

6.5 Distribution across cells of Normalised Fit: Pre, during and post balancing

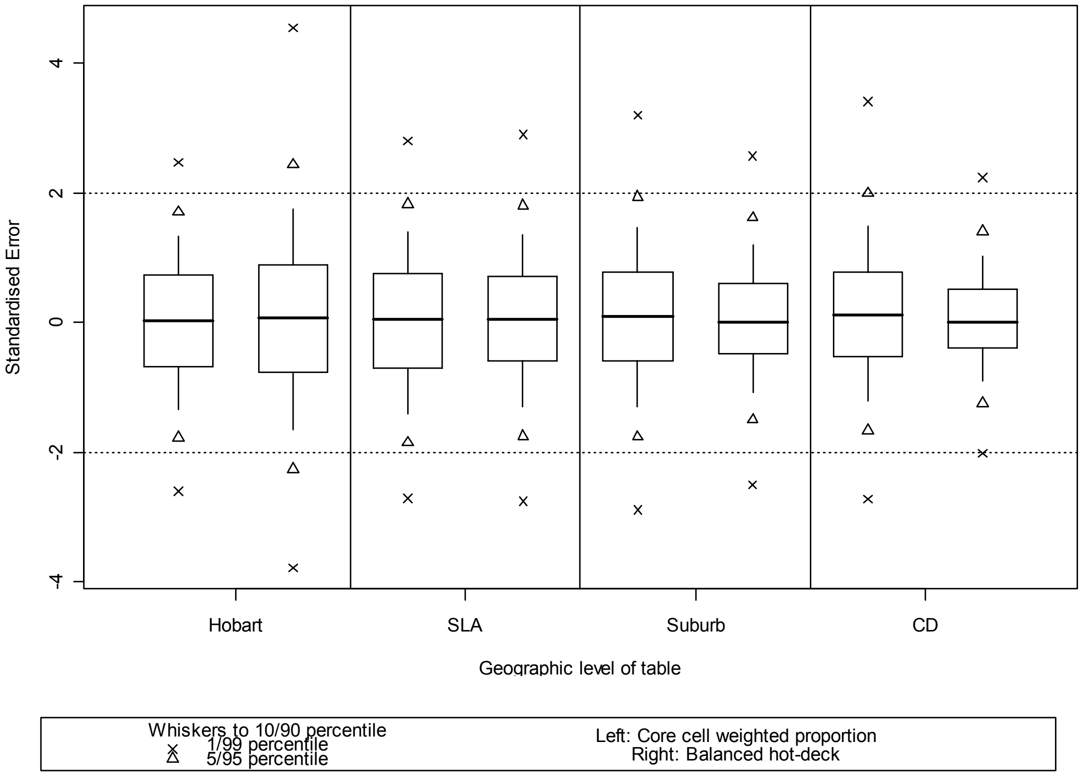


The fact the distribution of Normalised Fit is more spread for tables at broader geography is primarily due to the larger bias arising from the process of assigning donors. In addition, since the SE increases in proportion to the square root of the core cell value n_{gc}^{lq}, the Normalised Fit measure sets a tougher accuracy target for the cells at broader geography. Nonetheless graph 6.5 highlights how the balanced hot-deck does very well in greatly reducing the size of the bias resulting from the assignment of donors.

To illustrate the overall final performance of the balanced hot-deck, graph 6.6 compares the distribution of the Standardised Error for the balanced hot-deck and the ratio estimates $\hat{n}_{gct}^{lq}$. For cells at suburb and CD level, the discrepancy from the true cell value is on average smaller for balanced hot-deck than for the ratio estimates. At SLA level and all-of-Hobart a higher proportion of cells fit poorly to the true cell value for the balanced hot-deck compared with the ratio estimates.

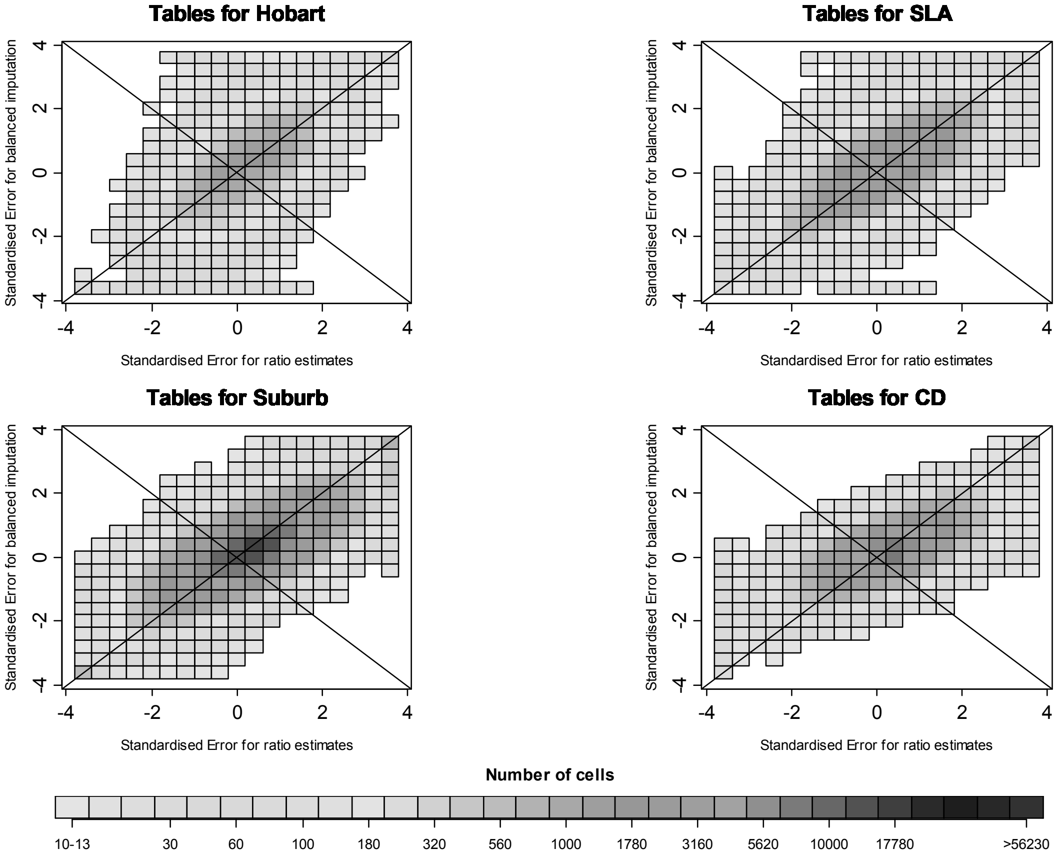
At first inspection, it appears remarkable that given the ratio estimates $\hat{n}_{gct}^{lq}$ provide the target cell values for the balancing, the balanced hot-deck is superior at fine geography. However, this result is evidence of the high sampling error on the ratio estimates at fine geography. The balanced hot-deck gives partially synthetic estimates for these small cells that are often closer to the true values than the ratio estimates.

6.6 Distributions across cells of Standardised Error for ratio estimates and balanced imputation



The results from graph 6.6 are explored further in graph 6.7, which presents a density plot of the Standardised Error for $\hat{n}_{gct}^{lq}$ (horizontal axis) against the Standardised Error for the balanced hot-deck (vertical axis). The plot area is divided into four triangles, with the regions within the left and right triangles indicating the Standardised Error is closer to zero for the balanced hot-deck compared to the $\hat{n}_{gct}^{lq}$. For Hobart and SLA level, a greater proportion of cells lie within the top and bottom triangles than the left and right triangles, indicating the error for the estimates $\hat{n}_{gct}^{lq}$ is on average smaller. At CD level there are more points in the left and right triangles, showing that at this level the balanced hot-deck provides superior estimates.

6.7 Density of Standardised Error for balanced hot-deck by Standardised Error for ratio estimate

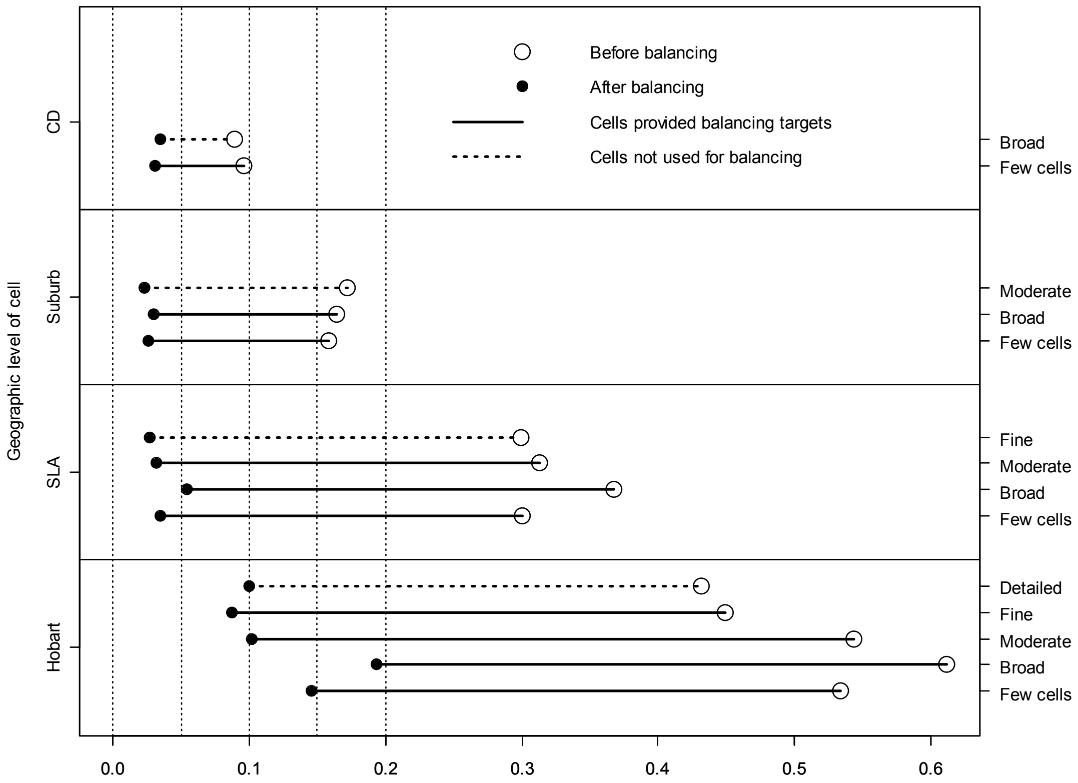


Considering the poor performance of the average hot-deck at broad geographic levels (graph 6.4) and the improvements obtained from the balancing process (graph 6.5), it seems reasonable to assume that the balanced hot-deck would produce very few poorly fitting cells if the implementation of the initial hot-deck imputation were more rigorous.

Impact on cells not used for balancing

Graph 6.8 compares before and after balancing the proportion of cells for which the Normalised Fit lies outside the range (–2,2). The comparison is broken down by the geographic level of the cell and by the Detail Level of the table to which the cell belongs. The cells represented by the dashed lines are 'evaluation only' cells (cells in Q^E but not Q) which did not provide a target value for balancing. Graph 6.8 shows the improvement in Normalised Fit of the evaluation cells is similar to the cells at the same geography which provided targets for balancing. The evaluation cells at SLA, Suburb and CD levels would have been assisted because corresponding cells at broader geography were involved in balancing. The evaluation cells for the all-of-Hobart tables did not have this assistance but still had enormous improvements in the fit measure. This outcome provides evidence that Census tables which are not used as constraints in the balancing process will not be adversely affected.

6.8 Proportion of cells with Normalised Fit outside (–2,2)



Improvements from balancing by detail of table

A question of interest is the types of cells for which balanced imputation performs best and worst. Graph 6.8 shows that:

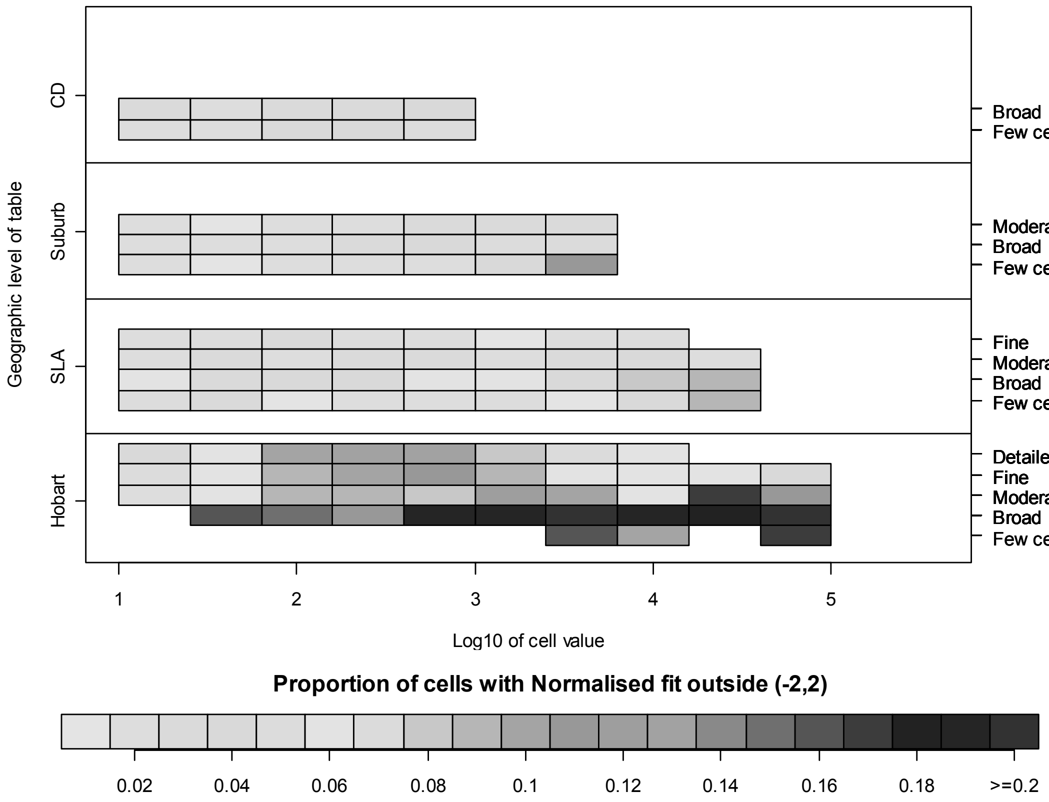
- for cells relating to all of Hobart, cells belonging to tables with fewer cells tended to have a higher proportion of poorly fitting cells before balancing.
- at each geographic level, the degree of improvement in Normalised Fit is very similar across the Detail Level.
- the extent of improvement is greatest for table cells at the broader geographic levels, which is unsurprising considering tables at these levels had more poorly fitting tables prior to balancing.

Analysis of balanced imputation by cell size

Graph 6.9 seeks to highlight relationships between the size of the table cell n_{gct}^{lq} and the likelihood of close agreement to the target value $\hat{n}_{gct}^{lq}$ after balancing. For tables at CD, Suburb and SLA level there is no strong relationship between accuracy and the value of the cell. The only cells at these geographies to stand out as poorly fitting are the largest cells in the Suburb and SLA tables which belong to the low detail tables.

For tables relating to all of Hobart there is more interesting behaviour, most notably that for the more detailed tables it is the mid-sized cells (100 – 1,300) which have less agreement with their target values.

6.9 Analysis of Normalised Fit by cell size



Analysis of balanced imputation by proportion of core class

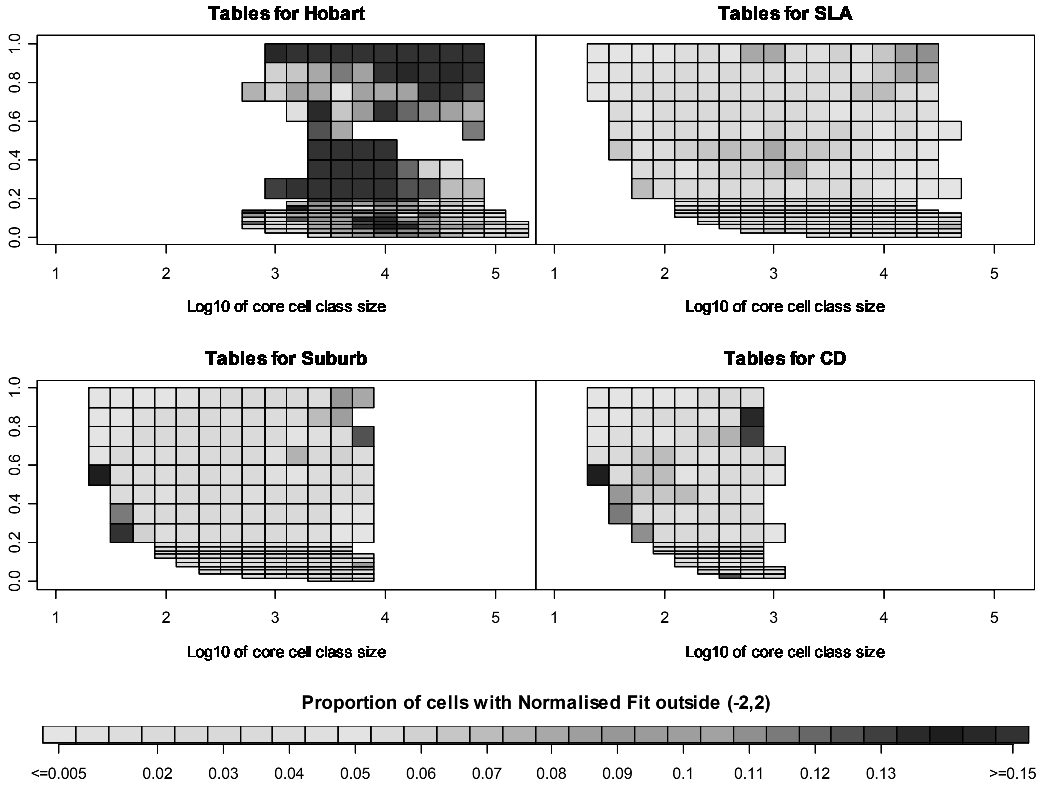
In the approximate SE formula (4), the SE of a cell value is driven by the size of the core cell n_{gc}^{lq} rather the size of the theme cell n_{gct}^{lq} itself. Graph 6.10 probes whether there is a relationship between quality of fit and the size of the core cell. In this graph the horizontal axis specifies $\log(n_{gc}^{lq})$, while the vertical axis is the proportion of n_{gc}^{lq} belonging to the theme class t for the cell. A consistent pattern in the plots is that the cells most likely to fit well are those with the combinations of:

- small n_{gc}^{lq} and high proportion belonging to the theme class t, or
- high n_{gc}^{lq} and low proportion belonging to the theme class t.

Quality of fit to the target values $\hat{n}_{gct}^{lq}$ is lowest for cells in which

- n_{gc}^{lq} is large and a high proportion belong to theme class t, or
- n_{gc}^{lq} is small and a small to moderate proportion belong to the theme class t.

6.10 Analysis of Normalised Fit by size of core class n_{gc}^{lq} and proportion belonging to theme class



6.3 Summary of findings

Overall, the balanced imputation approach has produced good outcomes as measured by the Standardised Error measure across the vast majority of table cells, including cells that were not used in the balancing. The fit measure is larger for tables at the broader levels of geography, however.

One reason for the larger fit values at broad geographic levels may be the use of the ratio estimates as target values for balancing. Ratio estimates for cells of different tables may be somewhat inconsistent with each other, making it impossible to achieve a close fit to all these cells. In a practical implementation this could be addressed when generating the target values.

It appears, though, that the cell totals obtained in this evaluation are close enough for most practical purposes, with most ‘poor’ fit values arising only because the standard error used as denominator in the fit measure is very small. If a realistic ‘required accuracy’ is defined for these cells then this could be substituted as a denominator for the fit measure, both in balancing and in measuring the outcomes.

7. DISCUSSION

7.1 Potential improvements to the balanced imputation algorithm

The algorithm described in Section 5 allows for imputation in a single pass of the data. If this does not achieve a reasonable fit to the ensemble of tables Q, there are some options for improving the imputation by performing multiple passes of the data.

One idea was suggested by Chauvet and Tillé (2006) in the context of balanced sampling. They note that better balancing can be achieved if the more difficult units are dealt with early, and easier units are left to the end. In the balanced imputation situation, this suggests that we note the units that had the largest impact on the fit measure during a first pass of the balanced imputation algorithm. These could then have their chosen imputes retained through a second pass of the data that is used to re-impute for the remaining, 'easier' units.

A second idea would be to conduct a number of passes of the data, and remove a single potential impute from consideration at each pass. The idea would be to obtain a more optimal outcome (in terms of lower Mean Squared Fit) by approaching the solution more gradually and thus making the outcome less dependent on order of imputation.

The quality of the final balanced product will suffer if there are major inadequacies in the hot-deck procedure that provided the potential donors. In the current study the hot-deck gave very poor results for persons reporting "Not stated" values for theme items. These persons will very often have many "Not stated" values among the core items as well, and using this in choosing potential donors would greatly improve the fit for cells involving "Not stated" values.

A related issue will arise if edit checks on the potential imputes are inadequate. This would be demonstrated as an impossible combination of a theme and core item value that does not appear in the reported theme data, but appears often in the imputed data. In this situation, balanced imputation will tend to reject the potential imputes that show this combination. By putting the onus for avoiding this combination onto the imputation algorithm, rather than applying an edit check beforehand, the available set of potential imputes is reduced. Overcoming this may require increasing markedly the number of potential imputes used as an input to the algorithm.

An analysis of the fit of the predicted imputes could be performed to identify this sort of problem before running the balancing algorithm. This may lead to identifying additional edits, after which the hot-deck imputation step would need to be re-run. A flexible tool like CANCEIS, which performs hot-deck imputation while taking account of edits, could assist greatly if this step is required.

7.2 Applications for balanced imputation in other contexts

Thematic forms provide a quite unusual setting for imputation, in that the missing values are for a probability sample of the population units. The usual imputation situation is very different – the mechanism that resulted in the data being missing is uncontrolled, and there is no inherent reason that the aggregate behaviour of the units reporting an item should apply to the units requiring that item to be imputed.

Nevertheless, balanced imputation does provide a mechanism for applying a richer model to the imputation of units without requiring an impossibly close match between donor and recipient units. For example, in a standard hot-deck it would be very limiting to classify the imputation groups by indigenous status to ensure that only indigenous donors are used for indigenous recipients. Balanced imputation allows an alternative mechanism for influencing the average imputed value assigned to the indigenous recipients.

Balanced imputation would be particularly useful in an imputation approach to adjusting Census counts for non-response. The idea here would be to impute additional persons and dwellings onto the Census file in such a way as to adjust for Census undercount and non-contact of units. In this situation, the Census post-enumeration survey (PES) provides good estimates at aggregate level of the Census night population by a variety of characteristics (Bell, Clarke and Whiting, 2007). Balanced imputation gives a method for controlling the imputation of additional dwellings and persons so that the result is consistent with these estimates.

7.3 Conclusions

The paper developed a balanced imputation methodology based on choosing from among a set of imputes provided by a hot-deck imputation procedure. The method was demonstrated using imputes from a hierarchical hot-deck at person level. The method would appear to be applicable to imputes provided by more complex imputation methodologies, such as the dwelling level imputation provided by the Statistics Canada package, CANCEIS.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Ray Chambers of the University of Wollongong for his role in reviewing an early draft of this paper, and for the number of helpful suggestions generated during discussions with him.

REFERENCES

Bankier, M. (2002) "2001 Canadian Census Weighting", *Proceedings of the Statistics Canada Symposium 2002*, Statistics Canada, Ottawa.

Bankier, M.; Poirer, P. and Lachance, M. (2001) "Efficient Methodology within the Canadian Census Edit and Imputation System (CANCEIS)", *ASA Proceedings of the Joint Statistical Meetings*, American Statistical Association.

Bell, P.A. (2000) "Weighting and Standard Error Estimation for ABS Household Surveys", Paper prepared for ABS Methodology Advisory Committee, July 2000, Australian Bureau of Statistics, Canberra.

Bell, P.A.; Clarke, C.F. and Whiting, J.P. (2007) "An Estimating Equation Approach to Census Coverage Adjustment", *Methodology Research Papers*, ABS cat. no. 1351.0.55.019, Australian Bureau of Statistics, Canberra.

Chauvet, G. and Tillé, Y. (2006) "A Fast Algorithm for Balanced Sampling", *Computational Statistics*, 21(1), pp. 53–62.

David, M.; Little, R.J.A.; Samuhel, M.E. and Triest, R.K. (1986) "Alternative Methods for CPS Income Imputation", *Journal of the American Statistical Association*, 81, pp. 29–41.

Deville, J.C. (2006) "Random Imputation Using Balanced Sampling", *Proceedings of the Q2006 European Conference on Quality in Survey Statistics*.

Deville, J.C. and Tillé, Y. (2004) "Efficient Balanced Sampling: The Cube Method", *Biometrika*, 91, pp. 893–912.

Pocock, S.J. and Simon, R. (1975) "Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trials", *Biometrics*, 31, pp. 103–115.

Statistics Canada (2002) *Sampling and Weighting, 2001 Census Technical Report*, Catalogue No. 92-395-XIE, Statistics Canada.

APPENDIXES

A. STANDARD ERROR OF ESTIMATED PROPORTION

Suppose that of the n_c people in a category c we have n_{ct} in a subcategory t, and the corresponding counts for the theme data are m_c and m_{ct}. The estimate of $p_{ct} = \frac{n_{ct}}{n_c}$ is then $\hat{p}_{ct} = \frac{m_{ct}}{m_c}$, which has

$$\mathrm{SE}\,(\hat{p}_{ct}) \doteq \sqrt{\frac{p_{ct}(1-p_{ct})}{n_c}\,\frac{(1-\pi)}{\pi}} \qquad (1)$$

<u>Proof</u>:

For $n_{ct}^{*} = n_c - n_{ct}$, $m_{ct} \sim$ Bernoulli (n_{ct}, π) and $m_{ct}^{*} = (m_c - m_{ct}) \sim$ Bernoulli (n_{ct}^{*}, π). These can be treated as independent. The Taylor Series linearisation approach gives

$$\begin{aligned}
\mathrm{var}(\hat{p}_{ct}) &\doteq \mathrm{var}(m_{ct})\left[\frac{\partial}{\partial m_{ct}}\left(\frac{m_{ct}}{m_{ct}+m_{ct}^{*}}\right)\Bigg|_{\substack{m_{ct}=\pi n_{ct}\\ m_{ct}^{*}=\pi n_{ct}^{*}}}\right]^2 + \\
&\quad \mathrm{var}(m_{ct}^{*})\left[\frac{\partial}{\partial m_{ct}^{*}}\left(\frac{m_{ct}}{m_{ct}+m_{ct}^{*}}\right)\Bigg|_{\substack{m_{ct}=\pi n_{ct}\\ m_{ct}^{*}=\pi n_{ct}^{*}}}\right]^2 \\
&\doteq n_{ct}\pi(1-\pi)\left[\frac{\pi n_{ct}^{*}}{(\pi n_{ct}+\pi n_{ct}^{*})^2}\right]^2 + n_{ct}^{*}\pi(1-\pi)\left[\frac{-\pi n_{ct}}{(\pi n_{ct}+\pi n_{ct}^{*})^2}\right]^2 \\
&\doteq n_{ct}\,\frac{(1-\pi)}{\pi}\left[\frac{n_{ct}^{*}}{n_c^2}\right]^2 + n_{ct}^{*}\,\frac{(1-\pi)}{\pi}\left[\frac{-n_{ct}}{n_c^2}\right]^2 \\
&\doteq \left[\frac{p_{ct}}{n_c}(1-p_{ct})^2 + \frac{(1-p_{ct})}{n_c}p_{ct}^2\right]\frac{(1-\pi)}{\pi} \\
&= \frac{p_{ct}(1-p_{ct})}{n_c}\,\frac{(1-\pi)}{\pi}
\end{aligned}$$

Q.E.D.

B. ITEMS USED FOR EVALUATION OF BALANCED IMPUTATION

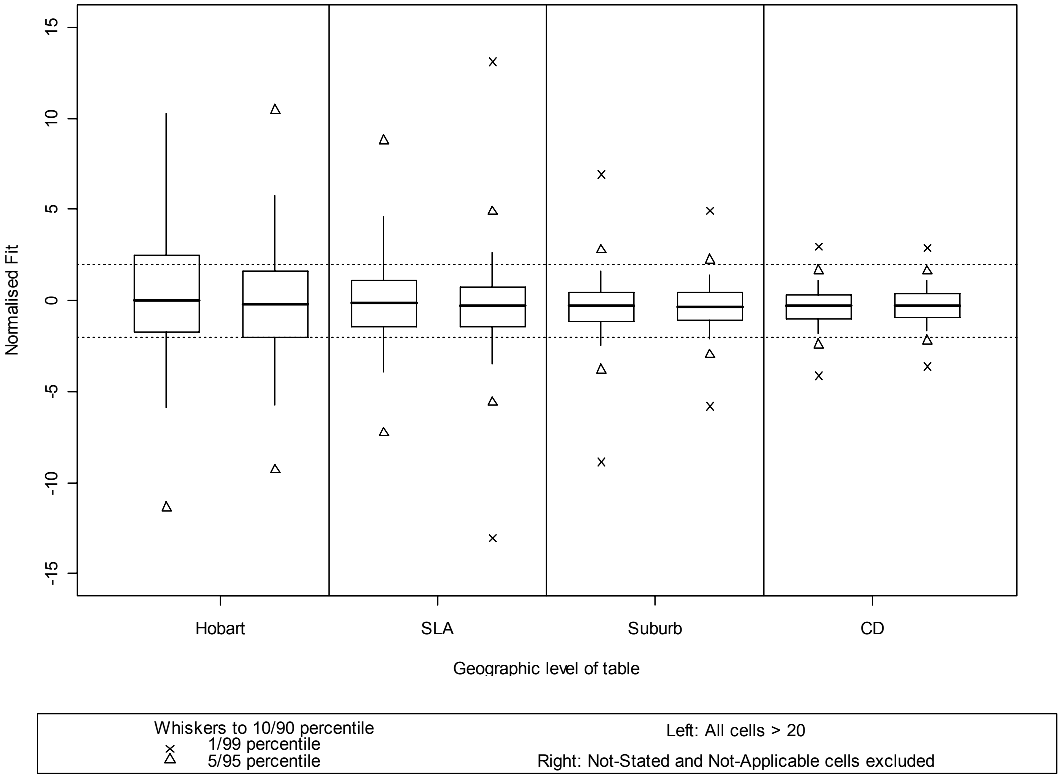
Table B.1 below lists the 2006 data items used as 'theme items' and 'core items' for the evaluation of the balanced imputation algorithm.

B.1 Theme and core items

	Number of variables controlling detail [number of classes taken by these variables]	
Core Items		
Sex	1	[2]
Age	4	[3, 7, 18, 13]
Role	1	[60]
Marital Status	2	[3, 4]
Indigenous Status	2	[3, 3]
Citizenship	1	[3]
Country of Birth	3	[3, 10, 36]
Highest Year of School Completed	2	[3, 7]
Highest Level of Educational Attainment	3	[3, 8, 19]
Education Status	4	[3, 5, 6, 12]
Labour Force Status	3	[3, 3, 5]
Hours Worked	1	[8]
Employment Type	2	[3, 4]
Type of Internet Connection in Dwelling	2	[3, 5]
Dwelling Structure	3	[3, 4, 9]
Household Composition	3	[3, 3, 6]
Language Spoken at Home	3	[2, 10, 59]
Year of Arrival	3	[4, 6, 23]
Theme Items		
Core Activity Need for Assistance	1	[3]
Unpaid Assistance to a Person with a Disability	1	[4]
Voluntary Work for an Organisation or Group	1	[4]
Industry of Employment	1	[21]
Individual Income	2	[4, 13]

C. FIT OF CELLS FEATURING "NOT STATED" AND "NOT APPLICABLE"

C.1 Normalised Fit for average hot-deck for all cells with $\hat{n}_{gct}^{lq} > 20$ and the set of cells which also exclude cells featuring "Not stated" or "Not applicable"



FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au

Produced by the Australian Bureau of Statistics