Small area/domain population statistics: when demand challenges fundamental estimation assumptions

Patrick Corr and Andrew Howe, Australian Bureau of Statistics

presented at the
IAOS ISI2005 satellite meeting:
Measuring Small and Indigenous Populations
Te Papa Tongarewa/Museum of New Zealand,
Wellington, New Zealand,
14 and 15 April 2005.

Abstract
The Australian Bureau of Statistics (ABS) prepares estimates of the resident population of Australia and its states and territories on a quarterly basis. These estimates are based on the results of the last Census of Population and Housing including adjustments for undercounting, residents temporarily overseas at the time of the census, and a range of administrative data on births, deaths, internal and international migration.

However, population estimates are also required for small geographic areas for a wide range of important applications: determining the distribution and relativity of grants to local governments, planning service delivery, electoral boundary redistribution and assessing government program performance and accountability at regional and small area levels. For many years the ABS has compiled estimates of the resident population for small geographic areas using a linear regression model with symptomatic indicators of growth as inputs. These are further disaggregated to age and sex, and areas comprising about 500 persons so that they can be re-aggregated into non-standard regions.

The evolution of administrative local areas into smaller population groups has challenged the assumptions and methods used to date. Increasing demand for finer level data for important government evaluations, decision making and research may exceed the capacity of these resident population estimates to support some of this precise fine level analysis. Small area data can be subject to third party scrutiny and criticism when the quality is perceived to be not consistent with that for national, state and territory population estimates or local expectations and program objectives. Also, small area resident population estimates may not always be the most appropriate estimate for some applications. This raises the question as to how national statistical Offices (NSOs) can respond to requests for fine level population data in an environment of increasing demand for evidence-based decision making, performance monitoring and evaluation at the local level. Examples of requests for fine level data and the challenges to statistical methods and assumptions are illustrated. Potential responses that NSOs may take are also discussed.
INTRODUCTION

This paper outlines the outputs and methods employed by the Australian Bureau of Statistics in compiling population estimates for small geographic areas and other small domains, such as age/sex breakdowns for sub-national regions, and Aboriginal and Torres Strait Islander Australians. Data sources used to derive these estimates, along with the various assumptions required and potential sources of error are also described, along with summaries of the accuracy of some estimates. A general conclusion is made that many small area population estimates, especially those with very small population cells, can be very inaccurate.

With increasing use of fine level population estimates for evidence-based policy and program decision making on a regular, sometimes annual, basis, the paper raises the question of whether variation in the resulting data is mainly reflecting the variability in the measurement and estimation errors present in the fine level data, rather than real differences in the population or indicators being assessed. Such over-precise analysis can cast a shadow of doubt over the quality of the overall dataset. However, the likely existence of the measurement and estimation imprecision could have been predicted in advance, even though it is unable to be quantified at the time the estimates are produced. Four real examples of such analysis are reported together with suggested actions national statistical offices can take when compiling and releasing small area and other small domain population estimates.

AN OVERVIEW OF AUSTRALIA’S POPULATION ESTIMATES

The Australian Bureau of Statistics (ABS) prepares Australia’s official estimates of the population and its components, based on the concept of place of usual residence.

Estimated Resident Populations (ERPs) are produced four times each year for Australia and its states and territories (which are from now referred to as ‘states’ in this paper), and once a year for Statistical Local Areas (SLAs). The SLA is Australia’s base spatial unit, which covers the entire country without gaps or overlaps. In 2001 there were about 1350 SLAs in Australia, each with an average population of around 14,000 persons. SLAs form, or combine to form, Local Government Areas (LGAs), which are legally designated areas over which local governments have responsibilities (ABS 2001a).

However, the SLA level of geography is not sufficiently detailed for many applications, for example: electoral boundary determination; postcode, suburb, electoral division or urban centre service delivery analysis; and other small community analysis including post-censal local government boundary change assessment. To address these requirements the ABS apportions these annual SLA level population estimates down to Collection District (CD) level. CDs are the smallest spatial units in Australia’s official geography and are also mutually exclusive and exhaustive areas. In 2001 there were about 37,200 CDs in Australia, each with an average population of around 500 persons. CDs are designed to facilitate delivery and collection of census forms, rather than for output purposes; however, because of their small size (and therefore their ability to be aggregated to client-specified regions), and relatively stable boundaries over time, CDs have become a popular output level in themselves.

The annual series of population estimates for SLAs are produced as at 30 June. The ABS first prepares SLA population totals, which are later disaggregated into age (single year of age up to 84 years, and 85 years and over) and sex components. The annual CD level estimates are also disaggregated by age and sex.

In addition to estimates of the total population, the ABS prepares experimental estimates of Australia’s Aboriginal and Torres Strait Islander (Indigenous) population in census years. For 30 June 2001 these were prepared concurrently as part of compiling the census year SLA level population estimates by age and sex, and re-aggregated to regional, state and national totals (ABS 2003a, ABS 2004). The ABS also apportions these SLA level Indigenous estimates down to CD level to enable aggregation to broader regions such as the former ATSIC regions and Electoral Divisions. The CD-apportioned estimates are not intended for release in their own right, mainly due to the high level of confidentiality required. Unlike post-censal estimates of the total population, shortcomings in the completeness and
consistency of reporting Indigenous origin status in the various administrative data make post-censal estimates non-viable, and therefore short range assumption-based projections compiled for the Indigenous populations of states and Australia are used (ABS 2004).

The ABS releases the total and Indigenous population estimates at the SLA level as standard products. The total population estimates by five year age group and sex are also released as standard output. However, estimates at the CD level are not standard ABS output – they require expert advice and are only released as a customised data service for specific client requests. Similarly, SLA estimates by single year of age are only released in response to specific requests on a consultancy basis. Because of the reduced status of these estimates, and the large amount of resources required to assess such a large number of cells, minimal scrutiny is applied to the CD level estimates, and to single year of age SLA level estimates; these estimates should therefore be treated with additional caution by users.

While the ABS prepares these population estimates based on where people usually live, many users of population data would prefer an indication of how many people are in regions at a particular point in time, for example in tourist areas, or daytime populations of business districts. Due to source data and conceptual definition limitations, the ABS (which refers to these populations as ‘service populations’) currently does not provide such estimates.

METHODS

Based on the data sources which are available in a timely manner, many different methods are used and assumptions applied to derive these population estimates, depending on:

- when the estimates are prepared for (census year, non-census year);
- what regions the estimates are prepared for (SLAs, CDs);
- what age/sex level the estimates are prepared for (all persons, age/sex breakdowns); and
- what type of population is being estimated (total, Indigenous).

Population estimates by SLA — census years

Australia’s Census of Population and Housing is conducted every five years. Census counts of usual residents are used to produce population estimates for SLAs, by age and sex, by:

Applying demographic adjustments to the census data. If a problem is discovered with census counts for a particular region (which had not been picked up and rectified previously in census validation) then they may be adjusted.

Adjusting for net census undercount. The ABS conducts a Post Enumeration Survey (PES) three weeks after the census to estimate the number of people (and their characteristics) who did not complete or were not included on a census form; it also detects the rarer instances of those counted more than once in the census. Net undercount is the excess of the undercount (people not counted) over the number of those overcounted (counted more than once). Incorporated in the 2001 census net undercount estimates is the effect of over-imputation of persons in dwellings believed to be occupied on census night but from which no completed return was obtained (ABS 2003b). Whilst net undercount is directly estimated for states, it is indirectly estimated for and applied to each SLA and the sub-national age by sex disaggregations.

Estimating the number of Australian residents temporarily overseas (RTOs). Derived from residential addresses reported on incoming passenger cards by residents returning to Australia after the census. A large sample of cards are coded to SLAs and weighted up to the total state level of RTO arrivals and added to the appropriate SLAs.

Reverting to the reference date. If the census does not occur on 30 June (the 2001 Census was held on 7 August) an adjustment is made to produce SLA estimates at the nearest 30 June reference date, using the component method (where the population at a point in time equals the population at a previous point in time, plus/minus the components of population change – births, deaths and net migration – which occurred between these two points in time).

3
**Indigenous population estimates**

In addition to these steps, census year population estimates at the SLA level are apportioned into Indigenous/non-Indigenous, based on responses to the census question on Indigenous status, with adjustments made to account for not-stated response to this question. The Indigenous estimates are regarded as ‘experimental’ due to the significant but unexplained intercensal growth in Indigenous census counts and the quality of the data on deaths (and to a lesser extent births) and migration data, which do not support the standard approach to population estimation (ABS 2004). This is especially the case at the SLA level. Consequently, Indigenous estimates are prepared at the SLA level for census years only (ABS 2003a).

**Total population estimates by SLA — non-census years**

Based on the census year population estimates, the SLA estimates are updated as at 30 June in following years. These are referred to as *preliminary* estimates. Although annual births and deaths data are available for SLAs, the absence of comprehensive and timely migration data for non-census years makes it currently not feasible to use the component method to prepare preliminary SLA population totals.

Instead, for most SLAs, a regression model is used to estimate the total population. The model establishes a relationship, based on the previous intercensal period, between the change in population and the change in symptomatic indicator variables. The indicator data must satisfy criteria regarding indicative ability (it needs to be a real indicator of population change), historical availability (it must be available and consistently defined over time) and geography (it should be available at the SLA level); it also needs to be available in a timely manner. The choice of indicators varies across regions and types of SLAs, and currently includes dwelling approvals (a proxy for dwelling completions) and Medicare enrolments (Medicare is Australia’s universal health care scheme). Using the models of best fit, and with the knowledge of the change in the indicators for the estimation period, population change since the previous census is estimated.

All modelled estimates are validated, and adjustments may be made to these modelled estimates after taking into account other indicator data, other methods and procedures, and/or advice provided by local experts. Some areas, for example where there are difficulties in obtaining reliable indicator data, or very small areas, are not modelled at all; and instead are estimated using other means – or may even have their census year estimates held constant over the intercensal period (ABS 2000).

**Age/sex population estimates by SLA — non-census years**

Once the post-censal total population estimates are derived for all SLAs they are disaggregated into age and sex. A variation of the component method is used to prepare these estimates, where the age/sex distribution of the population in a given year is updated from the previous year’s age/sex distribution, to which births are added, deaths are subtracted and net migration is added/subtracted.

In Australia, each birth and death is registered and coded to SLA. However, due to the absence of adequate migration data in non-census years, migration profiles derived from the previous census are applied to each SLA over the estimation period. To estimate the migration profile for each SLA in post-censal years, the relativities of the internal and overseas arrivals and departures, by age and sex, observed in the previous census are generally held constant over the estimation period for each SLA (ABS 2000).

All SLA population estimates are constrained to add to their (previously-derived) state level estimates. The SLA estimates are aggregated to form estimates for all other sub-state regions, such as local government areas, and capital city statistical divisions.

In a census year, preliminary estimates (updated from the previous census) are prepared. Later, when census results become available, *final* estimates for that year are also prepared. To overcome the break in continuity between the two data series (ie. preliminary non-census year and final census year estimates), all ERPs updated from the previous census are also recalculated to become final. In doing so, the difference as at the census year is assumed to have accumulated by an equal amount each year over the intercensal period.
Age/sex population estimates by CD

In the 2001 census, each respondent was coded to their CD of usual residence. To estimate census year CD populations, the census year SLA estimates, by age/sex, are apportioned into their CDs, based on the age/sex usual resident census count of each CD.

The total CD populations are adjusted for post-censal years to reflect any extra population implied by new building approvals since the previous census; the CD estimates are forced to add to the SLA level estimates, by age and sex.

MEASUREMENT AND ESTIMATION ERRORS

In attempting to determine the most accurate estimate of the population, in a reasonable time-frame, each stage of the estimation process outlined above is subject to errors.

Census counts

Census data is subject to a number of inaccuracies resulting from errors by respondents or in collection or processing. In the usual resident population counts the major potential sources of census errors are net undercount, partial non-response, processing error, imputation error (where records are imputed for persons believed to have been missed in the field phase of the census), random adjustment (where very small data cells are randomly adjusted or suppressed to avoid releasing information about particular individuals) and respondent error. Whilst many of these errors are corrected by careful processing and validation procedures, some may still remain and not be detected when compiling population estimates. The effect of the remaining errors is generally slight, although it may be more pronounced for smaller sub-groups of the population such as small regions, small age groups and the Indigenous population.

Post Enumeration Survey

For the 2001 census, the net undercount rate for Australia was 1.8%. Across the states, undercount was highest for Northern Territory (4.0%) and lowest for Australian Capital Territory (1.0%); young adult males had a higher undercount rate (3.8% for 25-29 year old males) than the remainder of the population; respondents of Indigenous origin had a higher undercount rate (6.1%) than the non-Indigenous population (ABS 2003b).

To arrive at these figures, the 2001 Post Enumeration Survey (PES) enumerated around 84,000 people, around 0.4% of the population. Despite its relatively large sample size (the PES is the largest household survey undertaken by the ABS), and the use of higher sampling fractions for smaller regions, the overall small sample size relative to the number of SLAs restricts the ability of the PES to directly measure undercount at SLA level – or for any small population. Net undercount is therefore synthetically estimated for each SLA, taking into account direct undercount estimates for broader region (state, capital city/rest of state), age, sex and Indigenous status.

The assumptions used in developing synthetic undercount estimates for SLAs cannot account for the geographic variability of true levels of net undercount if it could be perfectly measured. It is anticipated that true net undercount rates vary from SLA to SLA and from CD to CD in some form of distribution around the capital city/balance of state, territory and national rates.

While adjustment rates account for undercount differentials across region/age/sex/Indigenous status, sampling variation is generally higher for smaller populations. As a general rule, smaller population values have higher standard errors in the PES, and therefore the adjustments made to these smaller values (to derive the population estimates) can be assumed to be less reliable.

Residents Temporarily Overseas (RTOs)

Estimates of characteristics of RTOs by SLA at census time are derived from a sample of all incoming RTOs passenger cards completed on arrival in Australia, for which the residential addresses are coded to SLAs. The state age/sex distribution of RTOs is then applied to the adjustment to the SLAs within that state. While this is a large survey – approximately 30% (100,000) of all passenger cards in 2001 – this adjustment is also subject to sampling errors (in addition to non-sampling errors), with higher relative variation for smaller regions. Applying the state age/sex distribution of RTOs to SLAs may also be a source of error for those SLAs.
Reverting to the reference date

The adjustment from census date to reference date, using the component method, is also subject to potential inaccuracy. While the births and deaths which occurred between census date and 30 June can be reasonably accounted for (based on registrations data), migration can only be estimated.

Internal migration data is derived from the census question which asks for usual residence one year ago, which is coded to SLA. Migration for the period between census date and reference date is assumed to occur in proportion to the number of days between these two dates over the number of days in the year before the census. In newly developed or rapidly growing areas, this assumption may not be valid.

Indigenous population

In addition to the general census data errors mentioned earlier, particularly census net undercount, the census is unable to retrieve information on Indigenous status for a number of respondents due to non-response to the Indigenous origin question. Additionally, when a census form is imputed (when the collector cannot pick up a form from a person identified in the field) the Indigenous status is generally left as unknown. While adjustments for not-stated/non-response Indigenous status are made, by assuming the numbers of not-stateds/non-respondents are distributed according to the stated Indigenous status population (by SLA, age, sex and census-form type), the Indigenous population not captured in the census may not actually be distributed this way.

The Indigenous status of residents temporarily overseas on census night is also unknown, and is estimated on a pro-rata basis when preparing the population estimates.

Post-censal populations

The overall accuracy of post-censal SLA total populations has improved over time (Howe 2004). Even so, and while the SLA is usually an adequate level for the creation of population estimates of reasonable quality, the evolution of the SLA geography into smaller sized populations has led to some areas where the quality of data is questionable, mainly due to the small size of their populations and the assumptions required in their compilation. This aspect is much more apparent when the estimates for these areas are broken down into CDs, disaggregated into age/sex components, or apportioned based on Indigenous status.

As explained above, sub-state population change (total population) in Australia is estimated using the regression modelling approach, where models are determined based on the previous intercensal period (five years) and applied over the following intercensal period (five years). This assumes the relationship between the indicator variables and population is the same over both intercensal periods, a period of up to ten years. However, these relationships may change over this time.

In most current models the indicator variables used are dwelling approvals and Medicare enrolments – and each of these variables has its limitations. Potential problems with using dwelling approvals as an indicator of population change include approvals which don’t end in completion or are otherwise ultimately not occupied by usual residents (eg. holiday houses), and the timing of the lag between when a dwelling is approved and when it is occupied by usual residents (for each building approval, the models assume a six month lag between approval and people moving in); additionally, demolitions data is generally not available in Australia and therefore the direct effect of the removal of dwellings is not taken into account in these models.

The major quality issue with the Medicare data is that it is currently unavailable at the SLA of usual residence level; it is instead available by postcode – which is concorded (by the ABS) to SLA. Related to this issue is that for some areas, particularly in remote parts, the postcode for some Medicare enrolments reflects an Australia Post delivery route or a post office box which may be unable to be confidently concorded to SLA of usual resident of the person enrolled. Other potential data quality issues with Medicare include the lag associated with notifying a change in address, and the timeliness of taking deaths and overseas departures off the system.
For the age/sex breakdowns of these total populations, where the component method is used, problems may arise from not having a direct source of migration in to and out of SLAs. The age/sex estimation process assumes the age/sex structure of migration for each SLA is constant over the five years of the estimation phase, however this may well not be true, especially in small or rapidly changing SLAs.

The effects and consequences of SLA boundary changes are another potential source of error, due to the indicator data (for total estimates), and births, deaths and migration profiles (for age/sex estimates) being prepared on previously non-existent boundaries. When recalculating historical data on new boundaries, the relationship between the old and new boundaries is often assumed to be consistent over time (due to the absence of further information); this may not necessarily be the case, especially in rapidly growing regions.

It is assumed that the intercensal errors (described earlier) accumulate over time, and therefore the estimates further out from the base are less accurate than the earlier estimates. When populations are rebased using the results from the following census, these intercensal errors are apportioned equally over the intercensal period (due to the absence of any other information); however the error may have accrued at different rates throughout this time.

When the SLA populations by age/sex are apportioned down to CD level, additional errors may arise. Apportioning SLA populations down to CD based on the distribution of dwelling approvals (currently with no lag) is a simplistic approach – it does not and cannot directly take into account births, deaths or migration below SLA level. This method assumes that the combined effect of births, deaths and migration for each CD is reflected in the numbers of approvals for that CD.

**MEASURES OF ACCURACY**

Depending on what type of estimate is being assessed, some measures can give an indication of the level of confidence and/or accuracy in some of these estimates.

**Census year estimates — total population**

While undercount rates are applied to census counts by region/age/sex/Indigenous status – and therefore populations with a higher propensity to be missed in the census will have larger undercount rates applied to arrive at the population estimates – the undercount rates are all subject to survey sampling errors. For the total Australian net undercount rate of 1.8% in 2001, the standard error was a reasonably low 0.1 percentage points. By region, the standard errors of the undercount rates ranged from 0.2 to 0.7 percentage points. By age (five year age group)/sex, undercount rate standard errors ranged from 0.3 to 0.5 percentage points. Adjustments made for smaller populations have higher sampling variability, and therefore these adjustments are generally less reliable than adjustments made for larger populations.

**Census year estimates — Indigenous population**

Reflecting the relatively small Indigenous population, standard errors of undercount rates varied considerably by Indigenous status in 2001, from 0.1 percentage points for the non-Indigenous population to 1.1 for the Indigenous population (ABS 2002). The relatively high standard error of the Indigenous population compared to its undercount rate of 6.1% casts slightly greater unreliability on the Indigenous undercount adjustment compared to the total population.

The much larger number of census records for which there was no information on Indigenous status (767,757 – almost half of which are imputed records) compared to the total stated Indigenous count (410,003) also casts a revealing light on the accuracy of the total Indigenous ERP. The additional persons imputed from records with not stated Indigenous origin was 16,648.

In total, the 2001 census count of the stated Indigenous population (410,003) was effectively increased by 11.8% to arrive at the 30 June 2001 Indigenous ERP (458,520). In comparison, the total population census count (18,769,200) was effectively increased by just 3.4% to arrive at the total ERP (19,413,200).
Post-census year estimates — total population

In a census year, both preliminary estimates (updated from the previous census) and final estimates (derived from the current census) are prepared. The difference between the preliminary and final estimate for a census year is generally referred to as the intercensal error.

It is important to note that one or more of three aspects may contribute to an intercensal error: an error in the census-based population estimate in the previous census year (the base); an error in the census-based population estimate in the current census year (the end-point); and/or an error in the preliminary census year estimate (updated from the previous census year).

For the following analysis, the census-derived base and end-point estimates are implicitly assumed to be correct, and therefore the intercensal error is assumed to indicate the accuracy of the preliminary census year ERP (its method, application, data sources, etc) – and all preliminary ERPs prepared since the previous census. The census year intercensal error is assumed to have accumulated over the intercensal period, and therefore represents a ‘worse case’ scenario for these regions.

For Australia, the preliminary June 2001 ERP under-estimated the final population by 0.14% (–26,600 persons). For the states, the 2001 intercensal errors ranged from –1.6% (Australian Capital Territory) to +0.5% (Victoria) (ABS 2001b, ABS 2003c). At SLA level, the intercensal error ranged from zero to +267% (Mitchell, Australian Capital Territory: where the preliminary and final ERPs were 11 and 3 people respectively); considering populations of more than 1000 persons the largest intercensal error was +74% (Townsville City, Queensland); while for SLAs of at least 10,000 persons the largest error was –16% (Surfers Paradise, Queensland).

Summary statistics of the absolute values of these errors can be used to assess intercensal errors. The average absolute value of the intercensal errors for the 2001 SLA estimates (excluding SLAs with less than 500 persons) was 3.8%. Although this was a large decrease on the 1996 average (4.6%), errors for some regions remain high. Table 1 provides a summary of the average absolute intercensal errors for SLAs, by state, in 2001.

| TABLE 1: AVERAGE ABSOLUTE INTERCENSAL ERRORS, BY STATE, 30 JUNE 2001(a) |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
|                             | 1996 ('000 persons) | 2001 ('000 persons) | Number | Average size ('000 persons) | Median size ('000 persons) | Average % error (a) |
| NSW                         | 6,204.7             | 6,575.2             | 198    | 33.2                        | 11.9                        | 3.2               |
| Vic                         | 4,560.2             | 4,804.7             | 197    | 24.1                        | 12.0                        | 3.1               |
| Qld                         | 3,338.7             | 3,628.9             | 452    | 8.0                         | 5.7                         | 4.1               |
| SA                          | 1,474.3             | 1,511.7             | 122    | 12.2                        | 8.6                         | 2.4               |
| WA                          | 1,765.3             | 1,901.2             | 155    | 12.3                        | 2.9                         | 5.0               |
| Tas                         | 474.4               | 471.8               | 43     | 11.0                        | 5.7                         | 2.5               |
| NT                          | 181.8               | 197.8               | 64     | 3.1                         | 2.6                         | 7.2               |
| ACT                         | 308.3               | 319.3               | 106    | 3.0                         | 2.9                         | 3.8               |
| Australia (b)              | 18,310.7            | 19,413.2            | 1,340  | 14.5                        | 6.1                         | 3.8               |

(a) Average absolute intercensal error, excluding areas with ERP less than 500. (b) Includes Other Territories.

Average absolute intercensal errors for the 2001 series of SLA estimates increased with decreasing population size: SLAs with small populations recorded the highest percentage errors, while large SLAs recorded the lowest errors. Figure 1 provides an indication of the range of SLA sizes in 2001, and what proportion of the population was contained within these SLAs. For instance, SLAs with a population between 2000 and 5000 persons had an average absolute intercensal error of 4.1%; these SLAs comprised 27% of all SLAs in Australia, and contained 6% of the country’s population.

The relatively high errors for smaller populations may partly explain the relatively high average errors for SLAs in those states with lower overall SLA populations (Table 1).
Factors other than population size can also provide an overall indication of accuracy. A region’s population growth can also provide an idea of the size of the intercensal error for that region. Australia’s 2001 estimates showed that moderately growing populations were more likely to be estimated accurately than rapidly changing populations. Figure 2 shows the extent of average absolute intercensal error, based on population growth for SLAs between 1996 and 2001.

(a) Excludes SLAs with population less than 500.

Errors relating to growth may also partially explain the higher average errors for SLAs in those states with higher 1996 to 2001 population growth rates (Table 1): the states which grew fastest from 1996 to 2001 were Northern Territory (an 8.8% increase) and Queensland (8.7%); while Tasmania (–0.6%) and South Australia (2.5%) recorded the lowest population growth.

**Post-census year estimates — age/sex population**

To assist in gauging the accuracy of its sub-state age/sex estimates, the ABS calculated *provisional* estimates of the 2001 age/sex populations of SLAs. These followed on from the 1997 to 2000 series of preliminary SLA age/sex estimates. (Due to the impending release of final 2001 estimates, no 1996-based estimates of the 2001 age/sex populations of SLAs were released.) These provisional 2001 figures were therefore not scrutinised to the extent of the 1997 to 2000 preliminary SLA age/sex estimates. Nevertheless, an analysis of the...
provisional 2001 SLA age/sex estimates provides a relevant insight into the accuracy of the 1997 to 2000 preliminary age/sex estimates.

When the SLA estimates are broken down by age (five year groups) and sex, there are a large number of small cells. Based on the final 2001 estimates, 36% of SLA/age/sex cells had a population less than 100 persons. The median size of all SLA/age/sex cells was 162 persons.

Figure 3 indicates the accuracy of the 2001 provisional age/sex populations in SLAs by size of age/sex population cell. For instance, for age/sex cells of between 50 and 99 persons, which constitute 15% of all populated age/sex cells, the average absolute intercensal error was 17%.

The average absolute intercensal error for cells of between 10 and 19 persons (41% in percentage terms) equates to an average error, in numeric terms, of 6 persons.

Further analysis of the 2001 provisional age/sex SLA estimates reveals factors other than population size which affects overall errors. There also appears to be an age effect, as shown by Figure 4, which compares average absolute intercensal errors for the age/sex cells for all ages with errors for the population aged 20 to 24 and 25 to 29, by size of age/sex cell in 2001.

The relatively higher errors for 20 to 29 year olds reflects difficulties in estimating migration for these, the most transient of all ages.
DISCUSSION

Some ABS population estimates, for instance the finer-level disaggregations of sub-state estimates by age, sex and/or Indigenous status, or rapid growth areas, may be subject to large measurement or estimation errors. In particular, the very small cells – of which there are many – have the highest relative errors.

However, in contrast to the statistical measurement errors (sampling and non-sampling) in small area and small group population estimates, many of which are unknown at the time of estimation, the authors perceive increasing demand for fine level data for important government evaluations, program monitoring and decision making.

Examples of this use of fine-level data include:

- comparing population estimates by single year of age (5, 6, 7 and 8 year olds) for Indigenous and non-Indigenous by state with enrolments from the National Schools Statistics Collection enrolments to calculate school participation rates (SCRGSP 2003);
- comparing re-aggregations of CD-apportioned populations from SLA-level population totals for specific Indigenous communities and comparing these with composite estimates compiled from a range of administrative and other statistical information (Taylor and Bell 2002);
- monitoring school participation rates by single years of age and sub-state regions on an annual basis across a census rebasing revision period (in this example there was a large census rebasing revision to sub-state population estimates which resulted in a consequential revision to the past participation rates indicating that they had declined to a much lower level than first assessed using preliminary population estimates); and
- comparing aggregations of CD-apportioned ERPs for children aged 0 to 2 years up to Federal Electoral Divisions with number of childcare centre places (House of Representatives Official Hansard 2005).

Further, from the reaction of some stakeholders to small changes in population estimates from year to year, it is clear that an understanding of the precision limitations of the methods being used are not appreciated or well understood. While explanatory notes, caveats and quality statements are provided along with the estimates to users to explain these issues, the increased demand for finer-level population estimates by these users implies little appreciation of the extent of reliability of the small data cells, or of the capacity of the existing methods and data sources used to compile robust estimates for the purposes being employed. Rather, there is a general lack of appreciation of statistical measurement error amongst many key users of these statistics, and there is an expectation that all statistics produced have pristine accuracy.

How then might national statistical offices (NSOs) respond to the increasing demand for more precise estimates given the intended and past applications of the data? Several actions might be considered:

- ensure caveats and technical explanatory notes and an explanation of methods are easily accessible to users;
- seek out opportunities to inform and “educate” key users and stakeholders on the statistical limitations of small estimates and their fitness for purpose limits;
- model appropriate analytical behaviour – do not focus on large proportional movements in small cells;
- reconsider whether to compile and/or release data at excessively fine level disaggregations;
- be transparent with reporting of error analysis and other methodological investigations;
- develop and report quality or simple statistical confidence indicators when reporting estimates;
• institute procedures by which users of fine level data acknowledge that they understand the data quality limitations in advance of receiving the data;
• round all small area estimates, for example to the nearest 10, 50, 100, 500, to clearly convey the limitations of accuracy;
• acquire a wider range of indicators of population change and confront these data sources against each other and intercensal population change.

These are just some of the potential actions that could be taken by NSO’s when preparing and releasing population estimates. There may be others. The increasing demand for fine-level and more frequent population estimates for policy and program development, evaluation and reporting is presenting a tension between being responsive to client demands for data and a ‘duty of care’ responsibility to produce statistics which are accurate and fit for purpose.

The authors welcome discussion and comments on these proposals. Comments can be directed to andrew.howe@abs.gov.au or patrick.corr@abs.gov.au.

REFERENCES