Australian Bureau of Statistics

Demography Working Paper 2000/3

METHODS AND PROCEDURES FOR ESTIMATING SMALL AREA POPULATIONS IN AUSTRALIA

Andrew Howe
Small Area Population Unit
Australian Bureau of Statistics
GPO Box 2272 Adelaide 5001

andrew.howe@abs.gov.au

ABSTRACT

In Australia, population estimates for Statistical Local Areas are calculated annually. In years where a Census of Population and Housing is held, these estimates are derived with minimal adjustments from census data. These census–based population estimates are updated for following years based on estimation techniques which relate population change to change in indicators of population.

There are numerous methods available to update census–based populations. The Australian Bureau of Statistics currently relies on the regression estimation method to obtain a provisional estimate of the total population of each Statistical Local Area. Each population estimate is then validated individually using a range of indicator data, other techniques and utilising local knowledge.

ACKNOWLEDGMENTS

Many thanks to David Bartie, Megan Desira and Ken Holman for their many helpful suggestions and assistance in preparing this paper.

CONTENTS

1 INTRODUCTION

1.1 Overview

The Australian Bureau of Statistics (ABS) compiles and publishes estimates of the population and its components. Here, population is defined according to the concept of Estimated Resident Population (ERP), which links people to their place of usual residence within Australia. Population estimates are of fundamental importance to the community and receive specific mention in the Census and Statistics Act 1905.

Population estimates are produced annually, as at 30 June, for Statistical Local Areas (SLAs). The SLA is the base spatial unit used to collect and disseminate ABS statistics other than those collected from the Population Census. In aggregate, SLAs cover the whole of Australia without gaps or overlaps. SLAs conform to or combine to form Local Government Areas (LGAs).

The ABS first produces an estimate of the total population of each SLA. Later these totals are broken down into their age and sex components. This paper focuses on the methods used and procedures carried out to calculate the totals, and unless specified, the concept of a population estimate refers to the population total. The calculation of age/sex SLA estimates is summarised in appendix 1.

Unless otherwise specified, for the remainder of this paper the term 'state' refers to Australia's States and Territories.

1.2 Where small area population estimates are used

The estimated resident populations in SLAs are critical for state government grants bodies and local government authorities, especially in the allocation of funding and other resources. In addition, population estimates for SLAs are used as a base for population projections. ABS population surveys use SLA–based population estimates and projections as benchmarks. SLA population projections are also used by the Commonwealth and State electoral commissions. Other government agencies, health analysts, private enterprise and researchers also make extensive use of SLA population estimates and projections. This information is used for purposes such as planning the location of infrastructure, distributing resources, assessing needs and demands for products and services, and in the monitoring of social trends – especially in providing denominators for rates of indicators of the economy, health and the performance of public services.

1.3 Geography and population of Statistical Local Areas

SLAs are based on the boundaries of incorporated bodies of local government where these exist. These bodies are the Local Government Councils and the geographical areas which they administer are known as LGAs. In the Northern Territory, an incorporated administrative body gazetted under the Northern Territory Local Government Act can take the form of a Community Government Council. In the remainder of Australia where there is no incorporated body of local government, SLAs are defined to cover the unincorporated areas. In aggregate, SLAs cover the whole of Australia without gaps or overlaps.

In 1996, there were 1336 SLAs in Australia, including three which made up 'Other Territories' (Jervis Bay Territory, Territory of Christmas Island and Territory of Cocos (Keeling) Islands). These three territories are not included in this analysis, as their populations are estimated in the same way that state populations are estimated – using known migration data – rather than the SLA estimation procedure (ABS 2000).

All states excluding the Australian Capital Territory have an 'off–shore and migratory' SLA which are not spatial units in the usual sense – they have no defined boundaries. They are designed to facilitate the recording of people on census night who are off–shore on oil rigs, drilling platforms and other structures; on board vessels in and between Australian ports; or are in transit on board long distance trains, buses and aircraft.

In terms of both area and population, the sizes of SLAs vary substantially. The largest SLA in terms of area is Unincorporated Far North (South Australia), with an area of 670,376 square kilometres. This SLA comprises 68 per cent of the total land area of South Australia, or 9 per cent of the total area of Australia. The smallest SLA in terms of area is Narrows, located in Darwin, at 0.3 square kilometres.

In 1996, the populations of SLAs in Australia ranged from zero to well over 200,000 persons. The SLA with the largest population in 1996 was Blacktown (C), in the north–west of Sydney, with 239,818 persons. Several SLAs had zero population in 1996.

TABLE 1: DISTRIBUTION OF STATISTICAL LOCAL AREAS, BY POPULATION, 30 JUNE 1996

	Estimated Resident Population as at 30 June 1996							*Total number of SLAs*	*Mean ('000)*	*Median ('000)*
Number of SLAs	*0–2499*	*2500–4999*	*5000–9999*	*10000–24999*	*25000–49999*	*50000–99999*	*100000+*			
NSW	21	37	33	28	29	22	19	189	*32.8*	*10.9*
Vic.	10	34	48	39	45	22	2	200	*22.8*	*10.3*
Qld	98	110	130	101	7	3	0	449	*7.4*	*5.4*
SA	54	21	23	17	9	5	1	130	*11.3*	*3.5*
WA	73	18	18	19	13	9	1	151	*11.7*	*2.6*
Tas.	12	7	12	8	4	1	0	44	*10.8*	*5.9*
NT	33	21	8	1	0	0	0	63	*2.9*	*2.4*
ACT	41	54	11	1	0	0	0	107	*2.9*	*2.9*
Australia	342	302	283	214	107	62	23	1333	*13.7*	*4.0*

In 1996, all states except the Northern Territory and the Australian Capital Territory had a mean SLA population much larger than the median, indicating that in those states, a relatively small number of SLAs had a relatively large proportion of the state population. In fact, over 50 per cent of the population of South Australia and Western Australia resided in only 8 and 9 per cent of the SLAs respectively. At the other extreme, the Australian Capital Territory and the Northern Territory (where the median SLA size was much closer to the average SLA size) had half of their population in 27 and 25 per cent of SLAs respectively.

The growth patterns of SLAs also vary substantially by state.

Generally, states with a relatively high growth rate, and/or states with a large number of smaller SLAs – which are more susceptible to percentage population change – are more likely to have more rapidly growing and/or declining SLAs. Table 2 presents a summary of percentage population change for SLAs from 1991 to 1996.

TABLE 2: DISTRIBUTION OF STATISTICAL LOCAL AREAS, BY POPULATION GROWTH, 1991–96

	Change in Estimated Resident Population, 1991 to 1996								Total	*State*
Number of SLAs	*Below –10%*	*–10 to <–5%*	*–5 to <0%*	*0 to 5%*	*5 to 10%*	*10 to 20%*	*20 to 50%*	*50% & over*(a)	number of SLAs	*1991–96 increase (%)*
NSW	3	22	48	61	27	24	2	2	189	*5.2*
Vic.	3	24	60	58	26	18	7	4	200	*3.2*
Qld	24	53	76	69	56	72	67	32	449	*12.8*
SA	10	17	38	39	11	13	0	2	130	*1.9*
WA	25	26	29	20	15	25	8	3	151	*7.9*
Tas.	4	4	11	10	6	7	2	0	44	*1.7*
NT	5	9	6	13	4	14	9	3	63	*9.9*
ACT	23	31	18	9	3	7	5	11	107	*6.7*
Australia	97	186	286	279	148	180	100	57	1333	*5.9*

(a) SLAs which increased from a zero population in 1991 (this increase is mathematically undefined) are categorised as having 50%+ increase.

2 METHODS AVAILABLE TO ESTIMATE SMALL AREA POPULATIONS

Various methodologies can be used to obtain population estimates. One of the most critical aspects in estimating populations is whether data from a census can be used directly, or at least with minimal adaptations. As a major aim of a census is to enumerate populations, it should be used to produce estimates of the population wherever possible, particularly when the reference date of the estimates is at or close to the census date. However it is generally not satisfactory to directly use census–based population estimates for dates not close to a census.

2.1 Census years

SLA populations are required as at 30 June each year. Traditionally, the quinquennial Australian Census of Population and Housing is held at or near 30 June. So for those census years, census data – with minor adjustments – is used to estimate SLA populations. The method used in Australia is explained in greater detail in section 3.3.

2.2 Non–census years

Methods used to obtain small area population estimates updated from census–based estimates depend on the availability of data which can indicate the population change for these areas during the postcensal period. Ideally the population indicator data needs to be accurate, timely and available at the required geographic level.

2.2.1 Component method

The component method updates census–based populations by adding the births, subtracting the deaths, and adding the net migration which has occurred since the previous reference date. In symbolic terms:

$$P_{t+1} = P_t + b_{t,t+1} – d_{t,t+1} + m_{t,t+1}$$

for each area, where:

P_t = resident population of the area at time *t*

$b_{t,t+1}$ = births of residents of that area between time *t* and *t+1*

$d_{t,t+1}$ = deaths of residents of that area between time *t* and *t+1*

$m_{t,t+1}$ = net migration (arrivals to that area minus departures from that area) between time *t* and *t+1*.

The resultant updated population, P_{t+1}, can then be used as the base population for further updates, until the population can be re–enumerated based on a census.

This is the fundamental demographic equation, and is the ideal method of updating populations provided the components of births, deaths and net migration are available, timely, accurate and/or able to be estimated satisfactorily.

In Australia, births and deaths are available at the required SLA level (from data collected from the state registrars), however this data is generally not available in time to be used in the calculation of (preliminary) SLA population estimates. In addition, since residential movements do not need to be registered in Australia, net migration needs to be estimated by other means.

2.2.2 Regression methods

For the purposes of estimating small area populations, regression (or correlation) techniques first establish a relationship, based on past data, between population growth and the growth in 'symptomatic indicators', sometimes referred to as 'predictor variables'. Numbers of dwellings, births, deaths, drivers licenses and electricity connections are examples of these indicators. The relationships between population growth and symptomatic indicators are expressed mathematically in terms of regression coefficients and, with the knowledge of the growth in the indicators for the current time period, they enable population growth to be estimated. This is the method currently used to provisionally estimate SLA populations in Australia. Regression techniques are covered in more detail in section 5.

2.2.3 Dwellings–led methods

Dwellings, or housing–based methods, rely almost exclusively on updated or projected numbers of dwellings to estimate population.

The basis of this method is that virtually all of the population lives in some type of housing structure, whether it be a private or non–private dwelling. An area’s population can therefore be calculated by multiplying the number of occupied households by the average number of persons per household and adding the population resident in non–private dwellings.

This method assumes the number of households can be accurately enumerated or at least be satisfactorily estimated, the average occupancy ratio across all areas changes at approximately the same rate (or remains unchanged), and that the non–private dwelling population remains small, constant, or can be enumerated by other means.

An extension to this approach is to factor in changing (ie over time) and differential (ie over areas) occupancy rates for dwellings.

In Australia, most, if not all, states do not have the data to successfully implement the housing unit method at the SLA level. However its use should not be completely discounted, and, especially where accurate counts of dwellings can be obtained, results from housing unit methods can be relatively accurate.

2.2.4 Local census or survey

Generally expensive to run, but potentially rewarding, a local census provides a direct count of a small area population independent of any other source. Questionnaires can be sent to residents based on electoral registration lists. Non–response to these (generally non–compulsory) collections may be handled by taking into account results from a national (compulsory) census.

Sampling errors and even lower response rates than a census may discourage the use of a household survey to estimate populations. However, finer geographic and/or demographic characteristics may still be able to be estimated.

2.2.5 Other mathematical methods

Apportionment techniques break down the population of a large area into each small area, according to some indicator of each small area's population (the basic housing unit method is an example is this).

Composite methods derive estimates of age groups separately, which are then summed to secure a total for all ages.

Extrapolations may be made by observing the trend (linear or non–linear) between two previous points in time, and continuing this trend to the required reference date. It is more viable to consider this method for larger areas, which are less prone to significant variations in population change over time.

2.2.6 Human intervention

No matter what techniques are applied to obtain population estimates, each figure should be scrutinised by population analysts. Assessments should be made based on trends observed in a range of indicator data and/or local knowledge. Following these assessments, subjective adjustments may be made to the computed figures.

In some cases mathematical techniques should not be used, for example in areas where it is known that problems exist with indicator data. It may also be appropriate in particular areas to retain the census–based estimate during the postcensal period.

3 ABS METHODS AND PROCEDURES

3.1 History and overview of ABS small area population estimates

Annual population estimates for LGAs and/or SLAs have been published for Victoria since 1875, New South Wales and Queensland since 1911, South Australia since 1915, Tasmania since 1923, Western Australia since 1926, the Australian Capital Territory since 1968 and the Northern Territory since 1981.

Population estimates based on the concept of *usual residence* have been produced for LGAs and/or SLAs since 1976. Prior to this, LGA population figures were calculated on the *actual location* concept – that is, based on the number of persons actually present at that location at the given time (ABS 1983). Estimates based on actual location are generally higher in areas which attract short–term migrants, such as tourist areas.

SLA populations are calculated to add to the independently and previously derived state population estimates.

3.2 Timing and release of Statistical Local Area population estimates

Population estimates for SLAs are produced annually, as at 30 June. Generally, three series of SLA estimates are produced by the ABS. The three series tend to fall under the categories of preliminary, revised, and final. The release procedure depends on whether a census is conducted in that year. Since 1961, the Australian Bureau of Statistics has conducted a census of population and housing every five years.

3.2.1 Census years

In the interests of providing timely data, *preliminary* ERPs, updated from the previous census, are calculated. Later, when some census results become available, *revised* estimates for the census year are made. When final census results are available and the state estimates are finalised, *final* ERPs for SLAs are produced. Release dates for the 1996 Census–based population estimates are shown in table 3.

TABLE 3: TIMETABLE FOR RELEASE OF ESTIMATED RESIDENT POPULATION, STATISTICAL LOCAL AREAS, 30 JUNE 1996

Type	*Comment*	*Release date*
Preliminary	Updated from 1991 Census estimates	January – February 1997
Revised	Based on release of some 1996 Census data	July 1997
Final	Based on release of all 1996 Census data	December 1997

3.2.2 Non–census years

As is the case in census years, *preliminary* ERPs for SLAs are published about seven months after the reference date (preliminary state totals are released five and a third months after the reference date). When state totals are revised (about 15 months after the reference date), *revised* SLA totals are also calculated (to add to new state totals). This revision is usually made by apportioning the change in the state total across all SLAs (according to the size of their populations). However the adjustments to preliminary SLAs are, in most cases, minimal, because revisions to the state totals are generally very small.

Once the following census year ERPs have been finalised, and to overcome the break in continuity between the two data series (ie preliminary non–census year and final census year population estimates), all ERPs updated from the previous census are then recalculated to become *final*. In doing this update, it is generally assumed that the discrepancy as at the census year accumulates by an equal number each year over the intercensal period.

Population estimates for SLAs as at 30 June 1992 to 1995 were finalised in February 1998.

3.2.3 Age and sex components

SLA population totals are then disaggregated into age and sex components (see appendix 1). Preliminary SLA population estimates by age and sex are generally released within twelve months of the reference date.

3.3 Estimation procedures: census years

There are four major steps involved in the production of SLA population estimates from census data.

Step 1: persons coded to usual residence

Census counts of residents (ie excluding people counted in the census who usually reside overseas) by single year of age and sex are compiled for each SLA. The question in the 1996 Census from which this data was derived (question 7) was as follows:

> What is [each] person's usual address?
>
> - 'Usual' address is that address at which the person has lived or intends to live for a total of 6 months or more in 1996.
> - For persons who now have no usual address, write 'no usual address'.
> - For boarders at boarding school or college, give address at boarding school or college.

Step 2: adjust for census undercount

The census counts of usual residents by SLA are then adjusted to compensate for census net undercount using data from the Post Enumeration Survey (PES). The PES is a sample survey conducted immediately after the census to estimate the number of people (and their characteristics) who for one reason or another did not complete or were not included on a census form. The PES also detects instances of double counting of individuals, but the number of such cases is far outweighed by the number of people who are not counted. The net undercount is therefore the excess of the undercount (people not counted) over the number of instances of double counting. In 1996 the net undercount for Australia was 1.6 per cent (ABS 1997a).

The small sample size relative to the large number of SLAs restricts the reliability of the PES as a measure of undercount at the SLA level. Consequently, undercount is estimated using an iterative proportional fitting method (Purcell and Kish 1979). Based on the premise that undercount is related to age, sex and location, it is assumed that differentials for these characteristics at the SLA level reflect differentials at the state level for age and sex, and capital city/balance of state level for location.

The iterative proportional fitting method uses the following three data sets:
a) for each SLA, census counts of usual residents by single year of age and sex;
b) for states, census counts of usual residents by single year of age and sex, adjusted for undercount;
c) for capital cities/rest of state, census counts of usual residents by sex, adjusted for undercount.

Using the data sets in (b) and (c) as marginal controls, the census counts in (a) are adjusted in two stages:
(i) The first stage is to derive census counts of residents in capital cities/rest of state by single year of age and sex, adjusted for undercount. These are forced to add to both data sets (b) and (c) above.
(ii) The second stage is to adjust the census counts by age and sex for each SLA to match the estimates which were derived in the previous step. In all states except NT and ACT (due to their relatively small populations and therefore small sample size), this is done separately for SLAs in the capital city and balance of state.

Step 3: incorporating Australian residents temporarily overseas

Estimates of Australian residents temporarily overseas are then added to these adjusted census counts. These estimates are derived from data on residential addresses reported by these residents on returning to Australia after the census date, and are added to their respective SLAs. (All persons arriving in Australia, including Australian residents returning, are required to report their intended addresses which is taken to be their usual residence.)

Step 4: adjusting to 30 June

If the census does not occur on 30 June (for example, the 1996 Census was held 6 August) then a further adjustment is made to produce estimates at the nearest 30 June reference date. A variation of the component method (covered in section 2.2.1) is used for this adjustment.

Table 4 presents a historical account of the relationship between person counts obtained directly from the census and the population according to the definition applicable at the time.

TABLE 4: RELATIONSHIP BETWEEN CENSUS COUNTS AND SUB-STATE POPULATION, AUSTRALIA

Year	Adjustments made to obtain population
Up to 1966	Census count (30 June), actual location
1971, 1976(a)	Census count (30 June), actual location, adjusted for net census undercount
1981, 1986	Census count (30 June), usual residence (incorporating residents temporarily overseas), adjusted for net census undercount
1991–2001	Census count (August), usual residence (incorporating residents temporarily overseas), adjusted for net census undercount, adjusted to produce estimates at 30 June

(a) 1976 LGA estimates were later reconstructed based on the usual resident concept (incorporating residents temporarily overseas)

3.4 Estimation procedures: non–census years

Population estimates for non–census years are calculated annually for SLAs by updating the census–based estimates. A variety of procedures and methods have been used over the years to update census–based small area populations.

The methodologies and procedures applied by the ABS to produce SLA population estimates in non–census years have generally been performed separately for each state, although there has been a more centralised approach in recent years.

The different methods used can be broadly categorised into three time frames: before (and including) the 1981 census; between the 1981 and 1996 censuses; and beyond the 1996 census. Each of these periods are now summarised in turn.

3.4.1 Before 1982

Prior to 1982 most state offices of the ABS used various forms of the component model (see section 2.2.1) and/or methods based on projected growth trends, dwelling numbers, occupancy rates of dwellings and other similar information to derive a population estimate (actual location) for LGAs. Natural increase (births minus deaths) was calculated for LGAs and an attempt was made to estimate net migration for each LGA based on sources such as building data, occupancy ratios, school enrolments and electricity connections. 'Local knowledge' was also used where appropriate. These LGA estimates were then adjusted to sum to independently derived state totals.

While the 1981 estimates were generally considered acceptable in accuracy at the time, there were cases where cumulative errors over the intercensal period resulted in significant discrepancies when compared with census data. Around this time, the ABS curtailed its dwelling completions and demolitions collections – information that had been a key variable in the population estimation techniques until this time. This triggered the need to use different methods for population estimation.

Starting in the early 1970s, the ABS investigated alternative techniques to estimate LGA populations, especially the use of regression techniques (ABS 1977, ABS 1979). In 1982 the ABS released a working paper 'Regression Techniques for LGA Population Estimation' (ABS 1982), and in 1985 the Victorian office of the ABS released a paper extensively outlining the regression techniques and how it fared for LGAs in Victoria in the period 1976 to 1981 (ABS 1985). Regression techniques were briefly summarised in section 2.2.2, and are described in greater detail in section 5.

3.4.2 1982 to 1996

Most states had adopted regression techniques to estimate sub–state populations in time to produce the 1982 series of LGA/SLA population estimates. The following sections describe more comprehensively the methods and procedures used by each state from 1982 to 1996. The overall improvement in the accuracy of these estimates between 1981 and 1996 are summarised in appendix 2.

New South Wales

Population estimates were first calculated for Statistical Divisions (SDs), using two different methods. The first method was to analyse the proportion of population of the SD (to the state) for each year of the previous intercensal period and for each year since the previous census, and these proportions were extrapolated to the new reference date. The second method used regression techniques, using predictor variables such as dwellings, births, deaths and school enrolments.

Population growth for SDs were then evaluated in light of what was known about that SD (in terms of population mobility based on any recent internal migration surveys or growth in the indicators) and subjective judgements were made as to the most plausible growth for the SDs. Populations of the SDs were then forced to the state population.

SLA populations were also estimated using two different methods. The first looked at various data for each SLA, especially the change in the number of dwellings, changes in occupancy ratios over the past two intercensal periods and other indicators. Population growth for each SLA, however, was arrived at subjectively. The second method used regression techniques in a similar manner as was used for SDs. SLA populations were then forced to the SD populations.

Victoria

Population estimates were first calculated for Melbourne Statistical Division (MSD) and Rest of State (ROS). The MSD population was obtained by applying the ratio of the increase in the Victorian population to the increase in MSD dwellings during the previous intercensal period. The increase in MSD dwellings since the previous census multiplied by this ratio gave the increase in the MSD share of the Victorian population. The MSD population in a postcensal year was estimated by multiplying the change in the Victorian population since the previous census by the increase in population attributed to the MSD. This was added to the MSD population at the previous census. The ROS population was obtained by subtracting this from the state population.

SLAs in both the MSD and ROS were classified into groups (strata) according to the rate of growth in the number of occupied private dwellings in each SLA. Regression techniques were then used to estimate SLA populations. The symptomatic indicators used included occupied dwellings and Medicare enrolments. SLA populations were then forced to sum to the MSD and the ROS population estimates.

Queensland

Until 1986, SLA population estimates in Queensland were calculated from regression models. Symptomatic indicators included births, deaths, natural increase, school enrolments, dwelling stocks, number of electors, Family Allowance recipients and pensions, as well as the population total for the previous year. SD populations were also estimated from regression models using symptomatic indicators aggregated at the SD level, with the SLA population estimates forced to add to these totals.

From 1987 to 1996, the regression method continued to be used but a single model was derived for all SLAs (as opposed to a set of SLAs within a stratum as was used by some other states). In addition to the symptomatic indicators listed above, Medicare enrolments were also used. The growth in these symptomatic indicators during the previous year determined the SLA population estimate for the year being estimated. Each estimate was scrutinised in terms of its population growth and for other growth indicators. Finally, the selected population totals for all SLAs were forced to sum to the state population.

South Australia

Like Victoria, SLAs in South Australia were classified into low and high growth SLAs within Adelaide Statistical Division and ROS. For each stratum, the regression technique was used to link growth in the population variable with growth in several symptomatic indicators. Indicators included houses, 'other' dwellings, Family Allowance recipients, licensed drivers and Medicare enrolments.

For some SLAs, population estimates were obtained by a non–regression technique. This was done by relating the population in an SLA (as a percentage of the population of the stratum) with a similar proportion of a specific indicator, such as dwellings, Family Allowance recipients or licensed drivers.

A total population estimate for each stratum was made by summing population estimates for all SLAs (as opposed to obtaining separate estimates used as control totals in New South Wales or Victoria). An independent check of the population of each stratum was made using Medicare and other data. Population totals for all SLAs were then forced to sum to the state population.

Western Australia

Population estimates were first calculated for Perth Statistical Division (PSD) and ROS. For the PSD it was assumed that the rate of growth of the proportion of the state population resident in the PSD was the same as that for the previous intercensal period. This rate of population growth was then applied to the previous year's state population to estimate the PSD population. The ROS total was the difference between this and the population estimate for the state.

The regression technique was used independently for SLAs within PSD and ROS. Symptomatic indicators included dwellings, births, deaths, school enrolments and Family Allowance recipients.

Other relevant information was used to examine and supplement the estimates from the regression model, including information on significant local factors which might have caused population changes not picked up in the regression. Finally, all SLA totals within PSD and ROS were forced to sum to their separately calculated totals.

Tasmania

SLA population estimates were calculated in Tasmania using the regression technique. The symptomatic indicators used included births, deaths, school enrolments, building approvals, family allowances and the previous year's population.

The preliminary 1993 to 1996 SLA estimates relied less on mathematical techniques and were produced with the assistance of local government authorities and other bodies who provided information on significant local factors which may have caused population changes during the relevant period.

Northern Territory

Until 1985, sub–state population estimates were published only for the seven LGAs in the Northern Territory. However, from 1986, estimates were calculated for each SLA.

The total NT population estimate was apportioned to selected regional totals. Until 1991, the main indicator used for this was school enrolments. After 1991, changes in Family Allowance recipients and ABS building collections data was extensively used. Other indicators, including Medicare enrolments, electricity connections, births and deaths, and actual dwelling counts (for some areas), were also used to subjectively arrive at the population estimates.

Australian Capital Territory

Population estimates were produced for each suburb (SLA) of the ACT. The method used depended heavily on estimates of the number of occupied dwellings and their occupancy ratios. For each suburb the number of occupied private dwellings was estimated using domestic electricity connections and numbers of occupied dwellings supplied by the ACT Department of Environment, Land and Planning (or earlier from the National Capital Development Commission). Historical housing occupancy trends derived from census counts were used to estimate occupancy ratios for dwellings. These were then applied to the number of occupied private dwellings to estimate the population for each suburb. Births and student numbers were also used as indicators of population size.

3.4.3 1997 and beyond

In 1994 the ABS centralised some of the functions involved in the production of SLA population estimates, starting from the 1997 series of SLA total population estimates (and from the 1994 SLA age/sex estimates). Up until the 1996 series of preliminary SLA totals (ie those updated from the 1991 Census), each ABS Regional Office was almost entirely responsible for the estimation of SLA populations for their own state.

Reasons behind the decision to proceed with a more centralised approach included:

- a lack of backup support in state offices, whose resources generally enabled only one person (perhaps with the assistance of one other) to be responsible for the production of small area population estimates;
- a lack of training of new staff appointed to produce the estimates;
- a lack of documentation of procedures;
- a lack of a coordinated program of research and development necessary in the field of estimating populations of small areas;
- the repetition of state specific tasks that could be performed on a national scale, for example converting data to appropriate boundaries;
- a perceived demand for consistent nationwide small area data rather than the disjointed approach in place up until 1996;
- to adopt a more standard approach to SLA population estimation, based on a relatively robust method if available.

The creation of a core unit of several staff dedicated solely to estimating SLA populations was seen as an ideal way to address these issues. The centralised body – the Small Area Population Unit (SAPU) – is located in the Adelaide Office of the ABS.

Extensive testing in the ABS up to this time had concluded the regression approach was a relatively robust way of estimating SLA populations (see section 5). It is this method that has been primarily adopted by the SAPU.

It is important however to allow for local (state) control over these centrally–produced population estimates, which are obtained essentially from regression models. Hence each Regional Office of the ABS scrutinises the provisional estimates (ie those derived directly from the models), utilising other data sources and methods, and taking advantage of the essential aspect of local knowledge. With all provisional SLA estimates being calculated in a central location, each Regional Office is able to focus more on the population figures themselves. Repetitive and potentially time–consuming tasks, such as collecting and converting indicator data (which is now generally available on a national scale), and preparing and running computer programs, are performed centrally.

Each ABS Regional Office has the opportunity to adjust the populations provided directly from the regression models and/or to choose the most appropriate population figure if the results of more than one model has been provided to them. At this stage, the various State Government planning agencies have an opportunity to provide some input into the estimation process. Forums are held between the ABS and State Government planning agencies to discuss the provisional estimates for a selection of regions, particularly growing areas and those areas where estimation problems have been an issue in the past. This analysis and adjustment process is an essential aspect of the overall estimation procedure, and is primarily performed to account for population changes which are not (directly or indirectly) picked up by the regression models, for example a major residential expansion to a non-private dwelling. Section 10 outlines the current procedures undertaken by each ABS Regional Office after the provisional estimates (ie the estimates derived purely from the model) are obtained.

A similar procedure is followed for the production of the SLA population estimates by age and sex. The SAPU is responsible for the calculation of provisional estimates by age/sex (appendix 1) which are then forwarded to ABS Regional Offices for validation. In the validation process the Regional Office has the opportunity to query and/or adjust the age and/or sex components of the total population estimate.

The methods and procedures for estimating the SLA populations between 1997 and 2001 can be reviewed after the results of the 2001 Census of Population and Housing are made available. Preliminary June 2001 population estimates, based on updating the 1996 Census–based estimates and using the methods and procedures outlined in this paper, will be released early in 2002. Later in 2002, the first estimates derived from the 2001 Census will be made. A comparison of the preliminary (1996 Census–based) and final (2001 Census–based) June 2001 SLA estimates will give an indication of the performance of the current models. Section 8 gives an overview of the types of assessments that can be made between the two sets of population estimates.

4 METHODS AND PROCEDURES APPLIED BY OTHER NATIONAL STATISTICAL AGENCIES

The estimation methods and procedures of national statistical agencies in New Zealand, Canada, the United States and the United Kingdom are now briefly summarised. These countries are in a situation similar to Australia in that the migration component of population change has to be estimated. Some other countries, for instance those in Scandinavia, have in place a population register. With the inclusion of an address or some other locality indicator such registers reduce the requirement for components of population change – especially migration – to be estimated.

4.1 New Zealand

There is a census of population and housing conducted in New Zealand every five years. From the 1996 Census, estimates of small area populations were obtained by adjusting census counts, by usual residence, for net census undercount (an average of 1.2 per cent for the total New Zealand population) and for the number of New Zealand residents temporarily overseas on census night. This provided population estimates for small areas as at 6 March 1996. These figures were then updated for births, deaths and net permanent and long–term migration in each area between 6 March and the reference date of 30 June 1996.

Estimates of small area populations for years after the census are derived by updating the base (census–based) resident population in each area for estimated natural increase (births minus deaths), net permanent and long–term overseas migration and estimated internal migration. Symptomatic data are used to measure the internal migration. Due to the delays between the actual occurrence of births and deaths, and the registration of the event, it is necessary to estimate births and deaths to produce timely resident population estimates. As further information is received on births and deaths, the previously published estimates are revised, where necessary.

4.2 Canada

1996 Census data was adjusted for underenumeration, or 'Net Census Undercoverage'. This adjustment was based on the results of two studies: the 'reverse record check' (which provided information on persons missed in the census who should have been enumerated) and the 'overcoverage study' (which estimated the number of persons counted more than once, or counted when they were not part of the census universe). The 1996 Census data was also adjusted for incompletely enumerated Indian reserves, because a number of these reserves did not participate fully in the census (Statistics Canada estimated these populations separately). The Canadian census enumerates non–permanent residents – estimates of the number of these non–permanent residents are calculated and the census counts adjusted accordingly.

To estimate small area populations in non–census years, Statistics Canada uses both regression and component methods. The regression method is used to estimate Canadian census division populations by updating census–based populations according to changes in symptomatic indicators such as Family Allowance recipients aged 1–14 years, health insurance records and hydro connections. The component method updates the base population using the components of population change: births and deaths (obtained from the Canadian Centre for Health Information) and migration. Migration data is obtained from Revenue Canada tax files, and is broken down into immigration, emigration, interprovincial in–migration, interprovincial out–migration, intraprovincial in–migration and intraprovincial out–migration.

Preliminary census division populations are based on the regression approach and released first. Later, when the births, deaths and migration data is available, these population estimates are revised, essentially using the component method.

4.3 United States of America

Sub–national population estimates from the US Census (conducted every ten years) were not adjusted for net underenumeration.

Postcensal estimates are produced by the US Bureau of the Census for counties and places (cities, towns and townships). County estimates are developed with the component change method, which assumes that the components which constitute population change can be represented by administrative data series in a statistical model. Each component is estimated separately. For the household population, these components are births, deaths and net migration. For the non–household population change is represented by net change in the population residing in group quarters.

Each component used in the component change method is represented by data which are symptomatic of some aspect of population change. Registered births and deaths data are used – if births/deaths data for the current year are not available then they are estimated by using the previous years' data. Net movement from abroad is calculated separately, with an enumeration of legal immigrants and refugees (data from the US Immigration and Naturalization Service) and an estimate of the undocumented migrants made. Individual Federal income tax returns (not necessarily from the current year) are used to measure the internal component of migration, by matching the returns for successive years. Information from these tax returns is then used to derive a migration rate for each county, which is multiplied by the migration base (previous year population, plus births, minus deaths, plus immigration to/from abroad).

The non–household population is estimated separately using data from the Department of Defence and state–specific military barracks data. Additionally, numbers of college students living in dormitories, inmates of correctional and juvenile facilities and persons in health care facilities are estimated through an annual group quarters report submitted by state–based organisations.

County–level tabulations of the number of Medicare enrolees obtained from the Health Care Financing Administration are also used in the estimation of US county populations.

The US also has in place a Federal State Cooperative Program for Population Estimates, with all states participating. This program has as its immediate goal the development of population estimates for each county prepared by the state agencies designated by their governors to work with the US Census Bureau. Several methods are used by the states, with component, regression and housing unit–based methods widely applied.

4.4 United Kingdom

The Office for National Statistics (ONS) provides annually updated population estimates for the 459 local authority and 225 health authority districts in Britain (numbers as at 1996). The ONS adjusts census counts by: accounting for census underenumeration (about 2 per cent overall); coding students who are at their term–time address back to their home address; and taking into account the census date being different to the required mid–year reference date (the 1991 census was held 10 weeks before the 30 June reference date). These adjustments were made by ONS for local and health authority districts, but not for smaller areas.

Intercensal estimates are updated at annual intervals. To do this on a consistent basis, estimates are produced for 'building brick' areas. These are the intersections between local authority and health authority areas. Data on births and deaths are obtained from the compulsory civil registration system. The information on migration is less accurate – and only proxy indicators are available. Estimates of external migration are obtained from a sample of migrant contacts identified every year in the International Passenger Survey. The best available indicator of internal migration are the re–registrations on the National Health Service Central Register as people join doctors' lists in new areas after moving within the United Kingdom. These migration sources can only provide estimates down to the Family Health Services Authority (FHSA) level, which cover shire counties, metropolitan districts, or one or more London boroughs. To obtain estimates of migration at the smaller district level, data relating to changes on electoral registers are used to partition FHSA migration estimates to districts.

5 THE REGRESSION TECHNIQUE FOR ESTIMATING SMALL AREA POPULATIONS

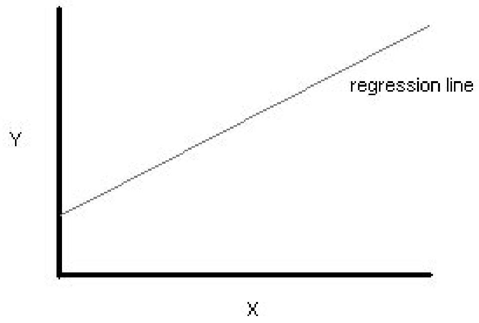
5.1 History

The regression approach for estimating populations was first proposed by Snow (1911) and reintroduced and modified by Schmitt and Crosetti (1954) as a means of estimating small area populations. Research papers such as those by Goldberg, Rao and Namboodiri (1964), Zitter and Shryock (1964), Namboodiri (1972), Ericksen (1974) and O'Hare (1976) were precursors to the ABS conducting its own research into regression techniques in the 1970s and 1980s. Results of this initial ABS research were presented in various publications (ABS 1977, 1982 and 1985).

5.2 Theory

Regression in its simplest form uses the relationship between two variables, one independent (X), and one dependent (Y). It assumes that there is a relationship between the two variables that can be expressed by a mathematical function. For our purposes we assume this mathematical function is linear. In sampling, this mathematical function will not be a perfect representation of the relationship between the variables as the values observed of the dependent variable (Y) are a random sample of Y values corresponding to a particular value of the independent variable (X).

We can show the relationship between the two variables graphically:



This means that given a fixed X value:

$Y = a + bX$ where a is the intercept of the line on the y–axis and b is the slope.

However, this does not take into account the random variation of an observed value given a particular X value. To take account of this random variation we use the following equation

$Y = a + bX + e$

where e is the error or random deviation of Y from the true value.

Expanding this to several X values, we obtain the equation:

$Y = a + b_1X_1 + b_2X_2 + ... + e$

where Y is the combined effect of several independent variables acting simultaneously. In our case the dependent variable is the population of the SLA, while the independent variables are the symptomatic indicators (dwellings growth, Medicare enrolments etc).

We can define the following:

P^{t}_{SLA} = SLA population at time t

P^{t}_{State} = State population at time t

$S^{t}_{i,SLA}$ = symptomatic indicator i for SLA at time t

$P^{t,t+n}_{SLA}$ = a measure of population growth in an SLA for the period t to t+n

$S^{t,t+n}_{i,SLA}$ = a measure of the growth in the symptomatic indicator i in an SLA for the period t to t+n.

For each SLA, the growth in both population and the symptomatic indicator is measured relative to the growth of the population or symptomatic indicator at the state level.

Regression techniques for population estimation are essentially based on two methods: the *difference correlation method* and the *ratio correlation method*.

In the difference correlation method, growth is measured in terms of changes in the differences between proportions of the state total in the SLA over time. Thus population growth between time t and t+n is expressed as:

$$P^{t,t+n}_{SLA} = \frac{P^{t+n}_{SLA}}{P^{t+n}_{State}} - \frac{P^{t}_{SLA}}{P^{t}_{State}}. \qquad (1)$$

Similarly, growth in the symptomatic indicator i is given by

$$S^{t,t+n}_{i,SLA} = \frac{S^{t+n}_{i,SLA}}{S^{t+n}_{i,State}} - \frac{S^{t}_{i,SLA}}{S^{t}_{i,State}}. \qquad (2)$$

So, regression analysis establishes the following relationship between growth rates of the population and the symptomatic indicator(s)

$$P^{t,t+n}_{SLA} = a + b_i S^{t,t+n}_{i,SLA} + e_{SLA} \qquad (3)$$

which is of the same form as the standard linear regression equation mentioned previously. The regression coefficients a and the b_i's are constants and their values are the same for all SLAs within a stratum, while e is the error term.

Times t and t+n refer to two consecutive census dates and allows us to derive values for a and b using the known values of the population and the symptomatic indicators at the dates of these two censuses (ideally the previous two censuses).

This allows us to estimate the change in population of an SLA from the last census to the reference date when the change in the symptomatic indicator for the SLA from the last census to the reference date is known. This can be expressed as:

$$P^{t+n,T}_{SLA} = a + \Sigma_i b_i S^{t+n,T}_{i,SLA}$$

where T is the reference date of estimation.

Substituting for $P^{t+n,T}_{SLA}$ from (1), we get:

$$\frac{P^{T}_{SLA}}{P^{T}_{State}} \qquad \frac{P^{t+n}_{SLA}}{P^{t+n}_{State}} = a + \Sigma_i b_i S^{t+n,T}_{i,SLA} \qquad \text{or,}$$

$$P^{T}_{SLA} = P^{T}_{State} * \left[\frac{P^{t+n}_{SLA}}{P^{t+n}_{State}} + (a + \Sigma_i b_i S^{t+n,T}_{i,SLA}) \right] \qquad (4)$$

The ratio correlation method is similar to the difference correlation method except that growth is measured in terms of ratios of proportions of the SLA variable to the state total rather than the differences of proportions. Thus (1) above is replaced with:

$$P^{t,t+n}_{SLA} = \frac{P^{t+n}_{SLA}}{P^{t+n}_{State}} \Big/ \frac{P^{t}_{SLA}}{P^{t}_{State}}$$

and equation (2) is replaced with

$$S^{t,t+n}_{i,SLA} = \frac{S^{t+n}_{i,SLA}}{S^{t+n}_{i,State}} \Big/ \frac{S^{t}_{i,SLA}}{S^{t}_{i,State}}.$$

The estimation equation is then given by:

$$P^{T}_{SLA} = P^{T}_{State} * \frac{P^{t+n}_{State}}{P^{t+n}_{SLA}} * \left[\frac{P^{t+n}_{SLA}}{P^{t+n}_{State}} + (a + \Sigma_i b_i S^{t+n,T}_{i,SLA}) \right]$$

Estimates from the difference correlation method automatically add to the sum of all SLAs unlike the estimates obtained from the ratio correlation method, which must be scaled to add to the state total.

Previous research conducted by the ABS (ABS 1982) established that the difference correlation method produced the most accurate SLA population estimates.

An example of an estimate made for a particular SLA population in 1998 is presented in appendix 3.

5.3 Stratification

Various research (Rosenberg 1968, ABS 1982) suggests that grouping areas into subsets which are more homogenous than the complete set is beneficial. More accurate estimates may be obtained using relationships which exist in these subsets. This aspect of stratification is used extensively in other fields of estimation such as sampling.

Several stratification strategies have been used in testing regression techniques to estimate Australian SLA populations. The most fruitful stratification techniques appear to be those based on population growth and geography. Other strata have been derived based on occupancy ratios, and the proportion of dwellings that are houses.

SLAs can be stratified based on growth derived from changes in the symptomatic indicators, for example change in dwellings.

Subsequent stratification of SLAs in to 'high' and 'low' growth may lead to greater accuracy in the estimates, since certain indicator variables, or linear combinations of variables, may be better suited to particular areas.

There must be a significant number of SLAs in each strata to enable a robust regression to be made.

Current ABS regression models use increase in dwellings since the previous census as the measure of growth. For each state, a wide range of thresholds for growth (numeric and percentage) were tested to determine which breakdown of high/low growth obtained the best results, based on the accuracy of the modelled 1996 estimates (see section 8).

Appendix 4 gives an indication of improvements to the modelled 1996 SLA population estimates when stratification is applied to the 1991–96 estimation models.

5.4 Stability over time

A major assumption of the regression techniques for estimating small area populations is that the relationship between the population and the indicator variables holds over the successive intercensal periods (for example, 1991–96 regression and 1997–2001 estimation periods).

Clearly, a comparison of coefficients between the previous intercensal and current intercensal periods cannot be made, since the coefficients for the current intercensal period regression are unknown until results of the next census are available.

An indication of the stability or otherwise of models and their coefficients may be made by comparing the regression coefficients between the two previous intercensal periods. Models with significantly inconsistent coefficients over successive intercensal periods are generally not considered for population estimation purposes.

5.5 Outliers

Plots, and the results of other tests of symptomatic indicators against population are analysed to detect and examine any apparent outliers in the regression models. Outliers are excluded from the regression to prevent the modelled relationship between population growth and symptomatic indicators being distorted.

5.6 Multicollinearity

Strong correlation between two or more explanatory variables in a regression model will result in high standard errors on estimates of the regression coefficients, and in the estimated coefficients being correlated rather than independent. This may affect the accuracy of the population estimates if the correlation pattern between the explanatory variables changes over time.

In the regression equations used to estimate SLA populations (see section 9), the inclusion of two or more highly correlated variables, for example variables which attempt to explain the same aspect of an area's population, is avoided. Examples of two highly correlated variables are Medicare enrolments aged 0–15 and Family Allowance recipients aged 0–15. Most variables used in the regression estimation equations are somewhat correlated, although to a lesser extent than the Medicare enrolments and Family Allowance recipients aged 0–15 example.

To determine the effect of multicollinearity, variables which are correlated with other variables can be deleted, and the effect of these deletions on the accuracy of the estimates can be analysed. This is a standard analysis performed when selecting the 'best' model from a range of models (see section 8).

6 EVALUATING INPUT DATA

To be suitable for the regression model, population indicator variables must satisfy criteria relating to availability, geography, consistency (of representation), and be indicative of change in population. Additionally, the variables used in the ABS regression models need to be available soon after the 30 June reference date to be incorporated into the preliminary SLA population estimates (due for release by February the following year).

These attributes are now examined in turn to draw up criteria for the ideal variable.

6.1 Availability

Variables used must be available for at least a ten year period, because the ABS models examine the relationship between population movement and variable movement over a five year intercensal period, and assumes the same relationship applies to the next five year period. For example, when producing intercensal estimates for the period 1997 to 2001, variables must be available from 1991 to 1996 – to estimate the coefficients – and 1996 to 2001 – when the models are being used to estimate population – making a ten year period from 1991 to 2001.

Ideally, the data should be available for fifteen years, to enable subsequent evaluation of its suitability for population estimation. To evaluate the performance of the indicators for the 1997–2001 estimation period, each variable should also be available for the model–determining period of 1986 to 1991, which can then be used to estimate 1992 to 1996 preliminary SLA populations. An assessment of the quality of the 1996 estimates can then be carried out by comparing with final 1996 (Census–based) SLA population estimates. Each indicator is assessed in turn. If it is established that the variable is a relatively good indicator of the 1996 SLA populations, the variable is then considered for the 1997 to 2001 estimation period.

6.2 Geography

Variables used must be available at the geographic level being modelled. Ideally the geography of the source should match the modelling geography, but sometimes this is not possible. Therefore variables which are capable of being converted to the modelling geography are also acceptable, as long as the conversion procedure is robust and consistent.

Currently in the ABS the modelling geography is at the SLA level and variable availability is at either SLA or postcode level depending on the source. Generally, data obtained from within the ABS is at SLA level, while data from outside the ABS is at postcode level.

6.2.1 Postcode to Statistical Local Area conversion

Concordances developed from census data are used to convert postcode level data to SLA level. The SAPU has investigated several postcode to SLA concordances for the purposes of converting data for use in population estimation.

The ABS has produced a standard postcode to SLA concordance for the 1991 and 1996 Censuses (ABS 1993, 1997b).

The 1991 standard ABS postcode to SLA concordance was produced using data from the 1991 Population Census Final Unit Record Files. That data included postcodes reported by respondents when filling out the Census form and SLA codes derived from the Census Collectors District (CD) in which the households were located. The SAPU scrutinised and adjusted the 1991 concordances, for example by investigating and reallocating some CDs, and parts of CDs, that were allocated to incorrect postcodes.

In 1996, the standard ABS Postal Area (postcode) to SLA concordance was derived by:

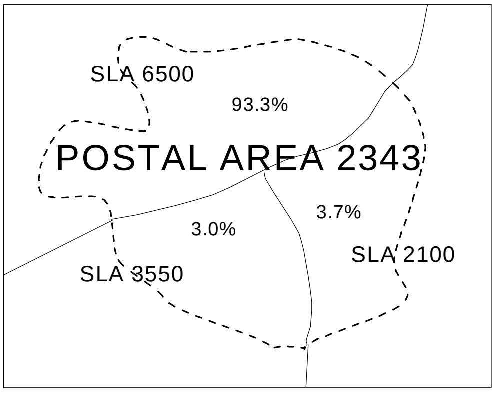
- obtaining a listing of CDs, their allocated postcodes, and their total population counts (in 1996, each CD was allocated to one postcode);
- aggregating CDs to SLAs, along with the respective postcodes and population counts;
- calculating the percentage of the postcode population which fell into each of the SLAs which intersected the postcode;
- validating codes against other sources of information.

Example records from the standard 1996 postcode to SLA concordance are presented in table 5.

TABLE 5: EXAMPLE RECORDS FROM THE 1996 STANDARD ABS POSTCODE TO SLA CONCORDANCE

Post-code	SLA code	SLA name	% of postcode population
2342	3550	Gunnedah (A)	28.0
2342	6300	Parry (A)	72.0
2343	3550	Gunnedah (A)	3.0
2343	6500	Quirindi (A)	93.3
2343	2100	Coonabaraban (A)	3.7
2344	6300	Parry (A)	100.0

FIGURE 1: POSTAL AREA 2343 IN THE 1996 STANDARD ABS POSTCODE TO SLA CONCORDANCE



(The standard definition of 'postcode' relates to Australia Post postcodes. In the ABS, 'Postal Area' refers to CD–derived postal areas, where each CD is allocated one postcode dependent on which Australia Post postcode contains the majority of the CD's population. Postal Areas are then constructed by grouping CDs with the same postcode. For consistency, the term postcode is also used to describe Postal Area in this paper.)

The 1996 standard ABS postcode to SLA concordance underwent a more rigorous verification process than the 1991 concordance. Hence, the SAPU adopted the 1996 standard ABS postcode to SLA concordance with minimal adjustments.

Another 1996 concordance was obtained by cross–tabulating the 1996 Census variables 'Postal Area of Usual Residence' and 'SLA of Usual Residence Census Night'. These variables were derived from question 7 in the 1996 Census (see section 3.2), which asked for each person's usual address, including postcode. The postcode variable is derived automatically for those enumerated at their place of usual residence from the postcode allocated to the CD on a best fit basis (for persons absent from their usual residence on census night, the postcode is coded using the address information provided by the respondent). The SLA variable is coded according to the National Localities Index, and in cases where a person does not state their place of usual residence, it is imputed (generally to place of enumeration). This concordance can be broken down into age and sex components by further cross–tabulating with the 'age last birthday' and 'sex' variables from the census.

Thus, four 1996 postcode to SLA concordances are produced:
a) the standard ABS 1991 concordance (adjusted by SAPU), updated annually to 1996;
b) the standard ABS 1996 census–based concordance;
c) census cross–tabulation of stated postcode of usual residence by SLA of usual residence; and
d) same as (c), but broken down by age/sex.

For each state, each concordance (a) to (d) was tested by converting the postcode level Medicare and Family Allowance data, and comparing the converted data to the census–based 1996 SLA population estimates.

Postcodes which do not pertain to a geographic entity, for example post office boxes, are generally omitted from the concordance. The exclusion of these postcodes is less of a problem when their base year postcode counts are of a similar size to the counts at the time of estimation – meaning the changes in share of state over the estimation period for the associated SLAs are less likely to be affected. If there does appear to be significant changes in the postcode counts then the SLA population figures need to be adjusted.

Since population estimates are produced each year, the concordances are updated annually during the intercensal period. The updating procedure accounts for postcode and SLA boundary changes, new or abolished postcodes and SLAs, and population growth differentials within postcodes. In updating the concordances, particular attention is paid to areas with boundary changes and high population growth.

6.2.2 Boundary changes

Data for areas involved in SLA boundary changes are backcast to the time of the previous census, based on the most recent boundaries. This is done to create a time series of indicator data on these boundaries. The converted indicator data can then assist in the calculation of historical populations on these updated boundaries.

For this reason it is ideal that the indicator variables be available at a lower level than SLA (for example Census Collection District), so that historical data for split SLAs can be calculated. This is done by building up the indicator data to adhere to the areas involved in the SLA boundary change, which can then be added/subtracted from the old SLA boundaries to correspond with the new boundaries.

Often it is not possible to obtain the data below the SLA level. In this case, a concordance (from 'old' to 'new' SLA boundaries) based on another variable (usually population) may instead be used.

To convert historical postcode level data, the postcode to SLA concordance for previous years must also be adjusted to accord with the updated SLA boundaries.

6.3 Consistency

The item represented by the indicator data must be consistently defined over the ten year regression and estimation period. Any change in definition of collection or coverage, processing procedures or timing can affect the population estimates.

Two major changes to variables which affected ABS regression models in the 1980s and 1990s were:

- Family Allowance recipients: means testing introduced from July 1987; and
- Medicare enrolments: a major purge of invalid records in the June quarter 1993.

If there is sufficient data available, these trend breaks can be backcast to the previous census date and adjustments made to the base data to reflect the 'new' definition or coverage of the variable. It may also be possible to do a transparent substitution of one variable for another, for example Medicare enrolments of children aged 0–15 years for Family Allowance recipients. An investigation of this particular example carried out by the ABS (ABS 1988) determined that transparent substitution appeared valid, based on data for South Australia.

Generally however, adjustments to the base data are made rather than substituting variables.

A change of reference date has never been considered a problem as far as growth indicator variables are concerned. Indeed, most administrative data is obtained at a date close to rather than at 30 June, a date which varies from year to year by as much as a fortnight. This variation is a small fraction of the year and has never been identified as a cause of error in population estimates.

6.4 Indicative ability

This refers to the ability of a variable to assist the model rather than hinder it. To be useful a variable must be able to indicate *real* changes since the previous census as they occur to the population. Therefore, ideally, the numeric value of a variable is a sizeable portion of the population (but not necessarily the total of the population).

The fact that indicators such as births, deaths and natural increase represent only a small proportion of the state's population infers that these variables are unstable indicators of population change. Changes in the value of these variables at SLA level are potentially amplified several times as the model applies them (along with other variables) to the change in the SLA's population. In other words, the model is too sensitive to variations in births, deaths and natural increase.

Another more intuitive argument against using births, deaths and natural increase in regression models is that the variation in the number of births, for example, in an SLA from year to year may well be explained by factors associated with an individual's or couple's choice (such as economic factors) rather than simply being a reflection of population change. And while it is true that rapidly growing SLAs will generate more births and deaths, there is generally an appreciable lag before the growth is displayed in the births and deaths. Thus the short term volatility generally swamps any long term trend of growth.

6.5 Timing

The ABS currently publishes the 30 June preliminary SLA population estimates by the following February. The 'provisional' estimates which are produced from the regression models, however, need to be available much earlier to incorporate enough time for scrutiny and analysis, any adjustments, preparation of commentary and formatting and preparation of publications. Once the indicator data is provided it must be checked for consistency, for example, by investigating any unexpected variations in the data. For data that needs to be converted to SLA level, time must also be set aside to carry out (and check) the conversion.

There are obvious time–saving benefits if the data is provided electronically, rather than the data having to be manually entered into the appropriate format.

6.6 Summary

To summarise, variables used for regression modelling of SLA populations should satisfy the following criteria:

- available for at least 10 years (preferably 15 years);
- available at the SLA level or capable of being converted to SLA level, on current and previous boundaries;
- consistent in timing, collection procedure and definition of coverage; and
- indicative of population change as it occurs.

Additionally, to ensure the timely release of population estimates, the data needs to be available soon after the date at which the population is being estimated. Electronic provision of the data is also vital.

7 VARIABLES USED IN ABS MODELS

The ABS currently uses, has used, and has investigated the use of numerous indicators of population change over the past twenty five years. These are listed at the end of this section in table 9.

Variables can fulfil two roles: as a base variable and/or growth indicator. The census supplies only base variables while ABS collections supply growth variables. Data sources from outside the ABS, for example Medicare enrolments and Family Allowance data, usually fulfil both roles.

Some variables can be combined to form a 'derived' variable (eg dwellings equals houses plus flats) while others can be split to better represent a particular section of the population (eg Medicare enrolments broken down into children aged 0–15 years, males aged 16 and over and females aged 16 and over).

Some variables cover the population better than others (eg family allowances over school enrolments) and some are better than others in detecting movements in the population (eg licensed drivers over school enrolments). Conversely, some variables are too volatile and movements within the variable do not necessarily represent movements within the total population (eg births, deaths).

The nature of some variables precludes them from indicating change of the whole population (eg school enrolments) while others exist only to complement others (eg flats). Some variables are duplicates of others or are so highly correlated (eg dwellings and houses) that they cannot all be used in the same model (see section 5.6).

Ideally each regression model should have variables which work together to indicate change in the whole of the population. Houses and flats together indirectly cover the whole population, however flats (and sometimes houses) alone are not sufficient to show population change. Factors such as changing occupancy ratios (persons per dwelling), non–private dwellings containing permanent residents and changing age/sex structures all combine to change the population beyond that indicated by building activity. The use of Medicare enrolments and Family Allowance recipients for children provides a direct measure of change in the population of children and an indirect measure of change in the population of their parents.

The following section provides a brief summary of variables which have been used and/or tested in ABS regression models. The ABS intends to test other variables over time should they meet (or come close to meeting) the criteria set out in section 6. An evaluation of data sources carried out by the ABS (ABS 1996), although directed at state–level estimation, discusses the potential usefulness of a range of other variables for estimating sub–state populations.

7.1 Variables currently used in ABS models

Currently the ABS regression models make extensive use of dwelling approvals and Medicare enrolment data, with Family Allowance recipients (Qld, SA, Tas, NT), drivers licenses (SA) and electricity connections (ACT) also used. Each of these variables is now discussed in turn.

7.1.1 Dwellings

Overall, the current regression models have dwellings as the most significant indicator of SLA population change. The ideal dwellings variable is the stock of *completed* dwellings, which are suitable for living and therefore most likely to contain usual residents.

Dwelling approvals are used as a proxy for dwelling commencements, or better still, completed dwellings. Latest available data from South Australia and Western Australia (ABS 1998a, 1998b) reveal that around 95 per cent of house approvals, and over 95 per cent of other residential dwelling approvals result in commencements.

The base for dwellings variables is the census count of dwellings, which is updated with dwellings approved since the census.

Dwelling approvals are compiled from: permits issued by local government authorities; contracts let or day labour work authorised by Commonwealth, State, semi–government and local government authorities; and major building activity in areas not subject to normal administrative approval (eg building on remote mine sites). Here a dwelling unit is defined as a self–contained suite of rooms, including cooking and bathing facilities and intended for long–term residential use.

Dwelling units can be created in one of four ways:

- through new work to create a residential building;
- through alteration/addition work to an existing residential building;
- through either new or alteration/addition work on a non–residential building; or
- through conversion of a non–residential building to a residential building.

The dwellings approvals data is satisfactory in terms of availability, geography, consistency and indicative nature (see section 6), and it is available on time. Dwelling approvals are geographically coded to CD level, which assists greatly in adjustments for SLA boundary changes (discussed in section 6.2.2).

For SLA population estimates, updated stock of dwellings are generally the most significant indicator of population growth, especially rapid growth. However, because the stock of dwellings is calculated by *adding* approvals to the count of dwellings as at the previous census, the number of dwellings cannot decrease. For this reason the SLA's share of state dwellings (rather than the number of dwellings) may be the better indicator of population change, and indeed it is the change in share of state dwellings that is used in the regression models.

Because of the time between the approval and completion of a dwelling ready for residential use, a six month lag for approvals is incorporated into the regression models. For example, the growth in population between 30 June 1997 and 30 June 1998 is assumed to be a result of dwellings approved between 31 December 1996 and 31 December 1997.

Houses

Approvals of dwellings can be broken down in to approvals of houses and other dwellings. Here a house is defined as a detached building primarily used for long–term residential purposes. It consists of one dwelling unit – for instance, a detached 'granny flat', or a detached dwelling unit (eg a caretaker's residence) associated with a non–residential building is defined as a house.

Sometimes by breaking down dwellings into houses and other dwellings, more accurate estimates are obtained. To use a house variable there must not be a significant number of other dwellings (eg flats) in the area, or at the very least the proportion of other dwellings remains relatively unchanged over time. Again, there is no direct indication of population decline with this variable.

Flats

Here flats is used to describe the buildings category 'flats, units or apartments', or 'other residential' dwellings. These are dwellings not having their own private grounds and usually sharing a common entrance, foyer or stairwell.

This data complements houses, and becomes significant for medium to high density housing areas. However the flats variable again has no direct indication of population decline, and it cannot stand alone as a variable (unless there is a high proportion of flats to total dwellings in the area).

7.1.2 Medicare enrolments

Since the early 1980s the ABS has received data on Medicare enrolments from the Health Insurance Commission. This data refers to enrolments at as at June each year at the postcode level, and is available within two months of the reference date.

While Medicare theoretically covers all Australian residents, there are some residents not covered due to access to alternative health sources, for example defence force personnel and some indigenous communities.

Medicare meets data criteria regarding availability, consistency, timeliness and coverage, however needs to be converted to SLA level using concordances.

As a variable in the regression model, *total* Medicare enrolments is generally not used. In theoretically covering the entire population, total Medicare enrolments tends to crowd out all other variables in these models. In other words, coefficients of other variables become insignificant. In this case, real population trends which aren't picked up by Medicare enrolments, and those areas which are not served adequately by the postcode to SLA concordance, have their estimation abilities diminished.

Experience has shown that Medicare variables become more useful in regression models when they are split into:

- children, aged 0–15 (which indirectly covers their parents);
- males, aged 16 and over; and
- females, aged 16 and over.

7.1.3 Family Allowance recipients, 0–15 year olds (Qld, SA, Tas, NT only)

Before the advent of means testing in 1987, this variable covered all children aged 0–15 years and indirectly covered their parents. Since 1987, Family Allowance (children) still covers the vast majority of the population aged 0–15 years (and indirectly covers their parents). However there could now be an income affect in that changes to Family Allowance recipients may reflect changing levels of income rather than population. In addition, this is a variable that needs to be converted from postcode to SLA level. Nevertheless this variable has proven to be a relatively reliable indicator of population change for some states.

7.1.4 Licensed drivers (SA only)

The variable gives a good coverage of the population aged 16 and over and benefits from the legal requirement for change of address to be notified as soon as possible following a move. However this data is only available at the postcode level, and hence needs to be converted from postcode to SLA.

7.1.5 Electricity connections (ACT only)

This data, provided by ACT Electricity and Water, is a good substitute for dwelling stock. However a consistent and long–term connections data series is only readily available for the ACT.

7.2 Variables previously used or tested in ABS models

Variables which have been used in the past for regression modelling, or have the potential to be used in future, are now discussed in turn. Strengths, weaknesses and, where relevant, reasons why variables are not currently used in the regression models are provided.

It must be stressed that this analysis considers only how these data sources behave in regression models. Even where a particular variable is found to be unsuitable for regression modelling, often the variable can be useful for scrutinising and checking SLA populations derived directly from the regression models.

Occupied houses

Like the 'houses' variable (which incorporates occupied and unoccupied houses), occupied houses are a most significant indicator of population growth. It might even be presumed that occupied houses are better in theory as an indicator of growth in SLAs with high numbers of holiday units. Numbers of occupied houses can be obtained from the census, with the occupancy rate held constant to estimate updated occupied houses from approvals data. However, like the 'houses' variable, 'occupied houses' is hampered by an inability to directly detect population decline, and the assumption of a constant occupancy rate is not always valid.

Testing has revealed no distinguishable improvement in quality of estimates when the occupied houses variable is used (when compared to all houses). During evaluation of the South Australian regression models, it was concluded that it was not worth the effort required to collect and maintain this variable.

Occupied flats

Advantages and disadvantages as discussed under 'occupied houses' are also applicable to this variable, which becomes significant for medium to high density areas. Occupied flats generally cannot stand alone as a variable (unless the stratum has a high proportion of flats).

Occupied dwellings

Obtained by adding occupied houses and occupied flats, this variable has the advantages and disadvantages as covered under 'occupied houses'.

Demolitions

Demolitions can be subtracted from dwelling approvals to theoretically obtain an accurate stock of dwelling. However, studies conducted for the Victoria and South Australia models have concluded that demolitions (estimated by electricity disconnections) made little if any impact on the quality of the populations estimates, and the cumbersome procedure of obtaining these figures was not considered worthwhile (the Victorian disconnections data used up until 1991 was obtained from several different sources in several different formats). This data is not readily available for other states.

Persons in dwellings

This variable converts indirect coverage of the whole population via dwellings data to direct coverage before being used in models. However, this: introduces another assumption; relies on constant (or constantly changing) occupancy ratios; and could crowd out other variables in regression models.

Births, deaths and natural increase

An essential part of the component model, births, deaths and the derived natural increase is available at the SLA level. But the use of one or more of these variables is limited in that they tend not to be a reliable indicator of total population change. Births are especially influenced by personal and economic factors, and deaths may be more dependent on the age structure of the SLA's population rather than being indicative of population change. Besides, the numbers of these components are generally too volatile (especially for small SLAs) to enable a robust regression to be made.

Births was used in the South Australia regression model from 1976 to 1986, and deaths was used from 1976 to 1981, however both data series were discontinued following evaluation of modelled estimates. Natural increase was tested after the 1986 census but quickly discarded.

Combined drivers licences and Medicare children, and combined drivers licences and Family Allowances

Both of these combined variables are representative of the entire population. Either of these variables tends to crowd out other useful variables. Both of these variables need to be converted from postcode to SLA. Both variables were previously used at some stage in South Australia regression models but have since been discarded.

Electoral enrolments

This provides a good coverage of population aged 18 and over, however electoral enrolments need to be converted from the source geography. In addition, this variable is not consistent in coverage, since it tends to peak just before an election.

School enrolments

This variable offers a potentially good coverage of children and, indirectly, their parents. However this data may be more specific to the area in which school is situated, rather than where students lived, and hence be unrepresentative of the resident population.

Motor vehicle registrations

This variable provides a good coverage of the population, however it is not a true growth indicator in that motor vehicle registrations may be affected by economic factors and hence the coverage may not be representative of true population distribution. This variable was used in the South Australia regression model 1981–86. However its volatility and the aspect of this variable competing with other variables as an indicator of change in the adult population led to its removal from the model.

7.3 Summary

Table 6 lists variables the ABS has investigated as possible indicator of population change in regression modelling over the past twenty five years.

TABLE 6: VARIABLES IN USE, OR PREVIOUSLY USED, BY THE ABS IN REGRESSION MODELS

TYPE OF POPULATION COVERED	TYPE			Used in
Variable	Base (census year)	Indicator	Derived	1997–2001 models
WHOLE POPULATION (INDIRECTLY)				
Houses, stock	●			●
Flats, stock	●			●
Dwellings, stock	●		●	●
Occupied houses, stock	●		●	
Occupied flats, stock	●		●	
Occupied dwellings, stock	●		●	
Electricity connections (ACT only)	●	●		●
Electricity disconnections		●		
House approvals		●		●
Flat approvals		●		●
House commencements		●		
Flat commencements		●		
House completions (NT only)		●		
Flat completions		●		
Persons in dwellings	●	●	●	
Motor vehicle registrations (SA only)	●	●		
Deaths	●	●		
Natural Increase	●	●	●	
WHOLE POPULATION (DIRECTLY)				
Medicare enrolments, all persons	●	●		
Combined licences & Medicare children	●	●	●	
Combined licences & Family Allowances	●	●	●	
ADULT POPULATION (AGED 16 AND OVER)				
Electoral enrolments	●	●		
Licensed drivers	●	●		●
Medicare enrolments, females aged 16+	●	●		●
Medicare enrolments, males aged 16+	●	●		●
Births	●	●		
CHILDREN (POPULATION AGED 15 AND UNDER)				
Medicare enrolments, children 0–15 years	●	●		●
Family allowances, children 0–15 years	●	●		●
School enrolments, primary only	●	●		
INTERNAL MIGRANTS				
Medicare movers – arrivals		●		
Medicare movers – departures		●		
OVERSEAS MIGRANTS (ARRIVALS)				
Medicare overseas arrivals		●		
N/A				
Occupancy rate (% occupied dwellings)	used to derive variables		n/a	
Occupancy ratio (persons/dwelling)	"		n/a	
Electricity consumption (ACT only)	"		n/a	

The variables in this table are sorted into groups according to the extent by which they represent the population. The coverage consists of four categories: whole population – indirectly; whole population – directly; adults 16 years and over; and children 0–15 years. Experience has shown that the better regression models contain a variable which *indirectly* represents the population as a whole, such as dwellings plus one or more variable(s) which *directly* represent specific age groups, such as child Family Allowance recipients. All regression models currently used to estimate SLA populations contain at least one variable from at least two of these categories. Note that some variables listed in this table have been discontinued or substituted with similar variables.

The table includes state specific variables.

There is the potential for other data sources to be used, or at least investigated for use, in the regression models depending on how well such variables meet the criteria set out in section 6.

8 ISSUES IN MODEL SELECTION

Section 5 indicates that there is a wide range of regression models to select from. In choosing the best model, there is a choice of the regression technique itself (difference or ratio correlation – see section 5.2). In stratifying SLAs in to more homogenous groups (section 5.3), there exists a wide range of stratification techniques and criteria (for example, the threshold values which distinguish between 'high' and 'low' growth SLAs). In assessing for stability of coefficients (section 5.4), outliers (section 5.5) and whether to exclude variables due to the possibility of multicollinearity (section 5.6), a multitude of alternative models can be considered. Aside from the types of models available, sections 6 and 7 indicate that there is a wide range of variables, and combinations of these variables, which can be used to develop the models.

The question then arises – from this extensive range of regression models, which is best?

8.1 Accuracy of estimates

The ABS considers a variety of aspects when choosing the 'best' model, however it is generally the accuracy of the modelled estimates that is the major consideration.

To assess the models used to estimate 1997 to 2001 SLA populations, regression models are first established based on data between the 1986 and 1991 censuses. The accuracy of the 1996 estimates derived from these models are then assessed. Once the best models have been established, separately, for each stratum in each state, the variables and their coefficients from the corresponding 1991 to 1996 regression models are held to estimate from 1997 to 2001, using updated indicator data.

To assess the accuracy of the modelled estimates, a comparison can be made between the predicted SLA estimates (from the regression model, based on updating the previous census–based estimates) and observed (or 'final', based on census) estimates. The difference between the modelled and final estimate is referred to as the *intercensal discrepancy*.

Several measures of the intercensal discrepancy can be considered. The most common measure is the *percentage* difference, however *numeric* differences and *logarithmic transformations* are also used. Table 7 expresses these measures mathematically, incorporating a simple example of each measure.

TABLE 7: MEASURES OF INTERCENSAL DISCREPANCY

Type of discrepancy	Formula P_m = modelled ERP P_f = final ERP	*Example, where:* *P_m = 10,500* *P_f = 10,000*
Percentage	$\frac{100 \times (P_m - P_f)}{P_f}$	*5%*
Numeric	$P_m - P_f$	*500 persons*
Log	$\frac{\log(P_m - P_f)}{\log(P_f)}$	*0.675*

There are advantages and disadvantages to each approach. Intercensal discrepancy may be dependent on the size of the populations being considered. There is a wide range of SLA sizes in Australia (see section 1.3). Evaluation based purely on percentage (or numeric) discrepancy assumes that each area's percentage (or numeric) discrepancy is as significant as the discrepancy in other SLAs, where it is most probable that, due to variations in size, this is not the case.

Given that SLAs range in size from zero to well over 200,000 persons, table 8 considers various measures of discrepancy associated with populations of size 2,000, 20,000 and 200,000.

TABLE 8: EXAMPLES OF DISCREPANCY AND MEASURES OF DISCREPANCY ASSOCIATED WITH POPULATIONS OF SIZE 2,000, 20,000 AND 200,000

Final ERP	Preliminary ERP	Percentage discrepancy	Numeric discrepancy	Log discrepancy
PERCENTAGE DISCREPANCY = +5 per cent				
2 000	2 100	*5.0*	100	0.61
20 000	21 000	*5.0*	1 000	0.70
200 000	210 000	*5.0*	10 000	0.75
NUMERIC DISCREPANCY = +1000 persons				
2 000	3 000	50.0	*1 000*	0.91
20 000	21 000	5.0	*1 000*	0.70
200 000	201 000	0.5	*1 000*	0.57
LOG DISCREPANCY = 0.70				
2 000	2 163	8.1	163	*0.67*
20 000	20 762	3.8	762	*0.67*
200 000	203 562	1.8	3 562	*0.67*

From this table it is apparent that the definition of an 'acceptable' discrepancy is not obvious. For a small area, a discrepancy of five per cent might be 'acceptable', for example a discrepancy of 100 for an SLA of 2,000 persons. However, a five per cent discrepancy in a region of 200,000 persons (10,000) is generally less 'acceptable'.

The percentage discrepancy is a straightforward concept and universally understood. However percentage discrepancies are not suited to analysis of SLAs which have significant differences in size, since the discrepancies are generally higher for small SLAs.

Analysis of numeric discrepancies assists in understanding magnitude of variation where per capita issues are important. Again however, evaluation purely based on numeric discrepancy is not suited to states with a large range of SLA populations. Numeric errors are generally higher for large SLAs.

A logarithmic transformation may be considered a 'compromise' between percentage and numeric discrepancy, because large SLAs are scaled downwards. However the concept of ‘log discrepancy’ is more difficult to understand, and an acceptable threshold is not obvious.

In selecting the best models from a wider range of models, there are two general approaches that can be used in assessing the intercensal discrepancies: selection of the models which produce the lowest average discrepancies across SLAs, or selection of the models which minimises the number of poorly estimated SLAs.

Table 9 presents an example of a (fictitious) range of population estimates (final and preliminary) based on three different models.

TABLE 9: POPULATION ESTIMATES AND SUMMARIES OF SELECTED MEASURES OF INTERCENSAL DISCREPANCY

Small area	**Estimated resident population**			
	Final	**Preliminary – derived from:**		
		Model A	**Model B**	**Model C**
Area A	350	310	360	325
Area B	1 200	1 270	1 180	1 020
Area C	4 260	4 110	4 170	4 300
Area D	8 860	9 880	9 440	9 520
Area E	22 470	22 140	22 710	23 060
Area F	24 330	23 130	25 130	26 170
Area G	64 120	63 920	60 920	63 130
Area H	65 020	66 370	60 030	65 390
Area I	110 670	109 680	115 490	108 310
Area J	198 720	199 770	199 670	198 580

SUMMARY MEASURES

	Model A	**Model B**	**Model C**
Average discrepancy	*Average (absolute) discrepancy (and rank)*		
Numeric (persons)	640 ('best')	1570 (3rd)	720 (2nd)
Percentage (per cent)	4.25 (2nd)	3.50 ('best')	4.50 (3rd)
'Log discrepancy' (‘log units’)	0.62 (3rd)	0.61 (2nd)	0.60 ('best')
Discrepancy greater than:	*Number of SLAs (and rank)*		
1000 persons	4 (3rd)	3 (2nd)	2 ('best')
5 per cent	3 (2nd)	2 ('best')	4 (3rd)
'Log discrepancy' > 0.67	2 ('best')	4 (3rd)	3 (2nd)

Discrepancies associated with models A, B and C are summarised under the 'average discrepancy' and 'discrepancy greater than:' sections of this table. The table shows that, based on the accuracy of the estimates, the selection of the 'best' model is not necessarily a straightforward decision.

In this example model A produces the lowest average numeric discrepancy (640 persons), model B the lowest percentage discrepancy (3.5 per cent) and model C the lowest 'log discrepancy' (0.6).

Based on minimising the number of 'poorly' estimated SLAs, the best model judged from having the least number of SLAs with: numeric discrepancy less than 1000 persons is model C (2 SLAs); percentage discrepancy less than five per cent is model B (2 SLAs); and log discrepancy less than 0.67 is model A (2 SLAs). The choice of these thresholds is subjective, and the thresholds chosen here are not necessarily standard.

ABS Demography Working Paper 98/1 (ABS 1998c) further describes details of the methods of evaluation of SLA population estimates and assesses the accuracy of the 1996 preliminary SLA estimates.

8.2 Other assessment factors

Stability of coefficients

As outlined in section 5.4, an indication of the stability of the regression coefficients may be gained by comparing the coefficients from regression runs from the two previous intercensal periods.

Inclusion of appropriate variables

Sometimes the average intercensal discrepancy for a stratum may be lowest when a particular variable is excluded – however commonsense would suggest that this variable should remain in the model, at least for some SLAs. This situation sometimes arises when the 'houses' (and not 'flats') variable is used in place of 'dwellings': particular SLAs may have a significant increase in flats over the estimation period, which would not be picked up in the model if the flats variable was not present. In this case it is advisable to have an alternative model to fall back on if the increase in flats is substantial.

Continued availability of indicator data

Alternative models should be considered if it is known that the supply of a particular data source will cease, or substantially lose its effectiveness to indicate population change, over the estimation phase.

8.3 Selection of 'best' models

After considering the issues outlined in sections 8.1 and 8.2, the 'best' models for each state and stratum within state are selected by ABS Regional Offices in conjunction with the SAPU. These models are selected at the beginning of the estimation (intercensal) period, and are generally assumed to hold for the remainder of this period. Models may be reselected later in the estimation period if more suitable models are subsequently determined, for example, if an indicator variable becomes inappropriate over time.

9 MODELS USED FOR ESTIMATING SLA POPULATIONS 1997 TO 2001

The criteria for classifying of SLAs into urban/rural, high/low growth and, for the ACT only, high/low flats are outlined in appendix 5. The following models show that urban areas tend to rely more on symptomatic indicators relating directly to people, such as Medicare enrolments instead of dwelling approvals. In contrast, rural areas show a greater reliance on dwelling approvals as a symptomatic indicator than urban areas do.

9.1 New South Wales

Stratum 1: urban, low growth SLAs
ERP = 0.41*dwelling approvals + 0.27*medicare children + 0.09*medicare women

Stratum 2: urban, high growth SLAs
ERP = 0.47*dwelling approvals + 0.20*medicare children + 0.30*medicare men

Stratum 3: rural, low growth SLAs
ERP = 0.97*dwelling approvals + 0.30*medicare children

Stratum 4: rural, high growth SLAs
ERP = 0.37*dwelling approvals + 0.34*medicare children + 0.22*medicare women

9.2 Victoria

Stratum 1: urban, low growth SLAs
ERP = 0.39*dwelling approvals + 0.25*medicare children + 0.29*medicare men

Stratum 2: urban, high growth SLAs
ERP = 0.51*dwelling approvals + 0.26*medicare children + 0.32*medicare men

Stratum 3: rural, low growth SLAs
ERP = *average of:* 1.08*dwelling approvals + 0.40*medicare men; *and*
1.05*dwelling approvals + 0.36*medicare children

Stratum 4: rural, high growth SLAs
ERP = 0.53*dwelling approvals + 0.38*medicare children

9.3 Queensland

Here we have two alternative and similarly performing models for each stratum.

Stratum 1: urban, low growth SLAs
ERP = 0.78*house approvals + 0.03*flat approvals + 0.16*medicare children
ERP = 0.84*dwelling approvals + 0.17*medicare children

Stratum 2: urban, high growth SLAs
ERP = 0.93*house approvals + 0.09*flat approvals + 0.06*family allowances
ERP = 1.00*dwelling approvals + 0.11*medicare children + 0.02*medicare women

Stratum 3: rural, low growth SLAs
ERP = 1.23*house approvals + 0.11*medicare children
ERP = 1.15*dwelling approvals + 0.13*medicare children

Stratum 4: rural, high growth SLAs
ERP = 0.75*house approvals + 0.04*flat approvals + 0.14*family allowances
ERP = 0.85*dwelling approvals + 0.15*family allowances

The Queensland regression models which break dwellings separately into houses and flats may provide a relatively low average discrepancy, but tend to have very small or zero coefficients for the 'flats' variables. Thus for SLAs with significant numbers of flats, the effect of flat approvals on the modelled population is small or nil. Therefore an alternative model which includes dwellings (combined houses and flats) is provided, and should be considered for those SLAs with a significant number of flats.

9.4 South Australia

Stratum 1: urban, low growth SLAs
ERP = 0.52*dwelling approvals + 0.46*medicare children + 0.36*drivers licences

Stratum 2: urban, high growth SLAs
ERP = 0.47*house approvals + 0.23*medicare children + 0.26*medicare women

Stratum 3: rural, low growth SLAs
ERP = 0.24*house appr. + 0.02*flat appr. + 0.17*family allow. + 0.26*medicare women

Stratum 4: rural, high growth SLAs
ERP = 0.66*house approvals + 0.22*flat approvals + 0.27*medicare children

9.5 Western Australia

Here we have three alternative models, which perform similarly, for each stratum.

Stratum 1: urban, low growth SLAs
ERP = 0.19*dwelling approvals + 0.83*medicare men
ERP = 0.49*house approvals + 0.06*flat approvals + 0.44*medicare men
ERP = 0.39*house approvals + 0.56*medicare men

Stratum 2: urban, high growth SLAs
ERP = 0.71*house approvals + 0.21*medicare men
ERP = 0.83*dwelling approvals + 0.23*medicare men
ERP = 0.85*dwelling approvals – 0.04*medicare children + 0.26*medicare women

Stratum 3: rural, low growth SLAs
ERP = 0.71*dwelling approvals + 0.64*medicare men
ERP = 0.69*dwelling approvals + 0.61*medicare children + 0.07*medicare women
ERP = 0.54*house approvals + 0.16*flat approvals + 0.64*medicare children

Stratum 4: rural, high growth SLAs
ERP = 0.77*house approvals + 0.07*flat approvals + 0.23*medicare men
ERP = 0.79*house approvals + 0.06*flat approvals + 0.15 medicare children
ERP = 0.99*dwelling approvals + 0.09*medicare children

9.6 Tasmania

Stratum 1: urban, low growth SLAs
ERP = 0.78*house approvals + 0.06*flat approvals + 0.12*family allowances

Stratum 2: urban, high growth SLAs
ERP = *average of:* 0.91*house appr. + 0.21*flat appr. + 0.09*family allowances; *and*
1.17*dwelling approvals + 0.07*family allowances

Stratum 3: rural, low growth SLAs
ERP = 1.14*dwelling approvals + 0.40*family allowances – 0.02*medicare females

Stratum 4: rural, high growth SLAs
ERP = 0.48*house approvals + 0.26*family allowances

9.7 Northern Territory

Stratum 1: urban, low growth SLAs

If growth in flats exceeds growth in houses:

ERP = 0.50*dwelling approvals + 0.27*medicare children;

else:

ERP = *average of:* 0.50*dwelling approvals + 0.27*medicare children; *and*
0.41*house approvals + 0.30*medicare children

Stratum 2: urban, high growth SLAs

ERP = 0.53*dwelling approvals + 0.33*family allowances

Stratum 3: rural, low growth SLAs

If growth in flats exceeds growth in houses:

ERP = 0.46*dwelling approvals + 0.46*medicare children;

else:

ERP = 0.44*housing approvals + 0.45*family allowances

Stratum 4: rural, high growth SLAs

If growth in flats exceeds growth in houses:

ERP = 0.54*dwelling approvals + 0.27*medicare children;

else:

ERP = 0.45*housing approvals + 0.24*family allowances

9.8 Australian Capital Territory

Stratum 1: high flats, high growth SLAs

ERP = 0.22*house approvals + 0.11*flat approvals + 0.34*medicare children

Stratum 2: high flats, low growth SLAs

ERP = *average of:* 0.68*dwelling approvals + 0.16*medicare children; *and*
0.59*electricity connections + 0.17*medicare children

Stratum 3: low flats SLAs

If dwelling growth is 'high':

ERP = *average of:* 0.86*dwelling approvals + 0.18*medicare children; *and*
0.80*electricity connections + 0.16*medicare children;

else:

ERP = 0.80*electricity connections + 0.16*medicare children

The populations of SLAs in the ACT which have a very high proportion of flats are estimated using the housing unit method (see section 2.2.3). Some SLAs in the ACT, due to their small size, have their census–based populations held constant in the post–censal period, unless there is an obvious indication of population change.

10 VALIDATING THE SLA POPULATION ESTIMATES

10.1 New South Wales

For each SLA, the modelled population growth since the previous census is compared with the annual population growth in the previous intercensal period. Areas where the regression model points to significantly different population growth in the previous year are especially scrutinised. Broader trends are also considered, for instance, throughout the 1980s and 1990s, large urban population areas in New South Wales (NSW) experienced positive growth while small population rural areas tended to experience declines in population.

Past analysis has found that dwelling approvals and total Medicare enrolments were positively correlated with population change, while family allowance recipients were weakly correlated. These results provide a basis to assess total population estimates for SLAs. Local knowledge and past trends are used to support or adjust the estimates.
A questionnaire is sent annually to all local government councils in NSW requesting information regarding the number of business rate notices, number of rateable residential properties, establishment or cessation of businesses, number of new or demolished residential dwellings and the council's estimate of total population – with an indication of the level of confidence placed in this figure. This data is compared with the data from previous years. The information may then be applied to the SLA estimates, taking into account the perceived relationship that exists between the change in businesses and number of dwellings and population change.

Extra attention is given to small SLAs, and SLAs with predominant 'special' populations such as mining or Indigenous populations. In these cases, the collection and analysis of SLA data is treated differently. For example, the council of the small LGA of Lord Howe Island provides the ABS annually with a head count of residents, including those temporarily away from the island. In SLAs with predominant mining communities, attention is given to the changing size of the mining workforce in these areas.

10.2 Victoria

Estimates produced directly by the regression models are scrutinised using graphs and by analysing the standard indicator data (especially Medicare and building approvals). Most of the estimates are considered acceptable at this stage with population growth being adequately explained by the indicators and supported by local knowledge. Other indicator data collected throughout the year from sources such as media reports and council notifications are potentially used to validate or adjust the modelled estimates. SLAs with major changes in population are scrutinised more closely and additional data sought, for example, from individual councils.

Some changes are made to modelled data to account for: prisoners but not prison officers (from the Office of Corrections) — only where new prisons have opened or old prisons have closed (continuing prisons are assumed to have relatively constant numbers); military personnel; the (reduced) effect of new dwellings in holiday resort areas eg Alpine (S); and other effects that cannot be immediately justified.

In 1998, for the first time, a letter was sent to representatives of all LGAs in Victoria offering them the chance to notify the ABS of anything which might impact on the population of the LGA. Response was purely voluntary and there was no follow-up of non–responses. Prior to 1998 there had been occasional approaches from LGAs eager to supply information. Recent information provided to the ABS included details of a new prison (the details of which were also obtained through the Office of Corrections) and details of an increase in the numbers of personnel at a military establishment (Puckapunyal Army Camp). It is proposed to continue to send this letter annually.

10.3 Queensland

Individual SLAs show particular characteristic relationships between the regression model's indicator variables and population growth. To help understand these relationships at the individual SLA level, Queensland analyses time–series graphs for Statistical Subdivisions (SSDs) and Statistical Divisions (SDs) in addition to the SLAs.

Firstly, for all SLAs, recent population growth as displayed by the regression models are compared with growth observed in the previous intercensal period. If the modelled estimate seems to continue the growth observed in the previous intercensal period, and other information doesn't indicate changes (for example new residential developments not adequately quantified by building approvals data), the modelled estimate generally remains unaltered.

SLAs with more unusual patterns are given closer attention. Their modelled populations are reviewed against indicator data not already used in the model, utilising information from sources such as: individual councils; the Queensland Corrective Services Commission Annual Report; school enrolments; the Australian Electoral Commission; census data; nursing home bed numbers; registered births and deaths rates for previous years; notes from the previous year's estimation process; and information from the media, real estate publications etc. Modelled estimates may then be adjusted up or down by a certain percentage or to a particular figure.

Time series graphs for each SD and SSD are viewed to check consistency with trends from the previous intercensal period in a process similar to that described for SLAs. For areas showing peculiar patterns, or patterns which are inconsistent with expectations, SLAs within the area are reviewed. It may then be necessary to alter the populations of individual SLAs, or all SLAs within the larger area to force to a new SD or SSD total.

All figures are then checked again. Particular groupings of areas with certain characteristics are especially re–examined. A different person then reviews each adjusted SLA, SD and SSD figure.

10.4 South Australia

Modelled population growth during the estimation period is considered in relation to indicator data growth. Typically this includes counts of Medicare enrolments, dwelling approvals, drivers licences, and family allowance recipients. Of particular concern is how changes in population over time relate to indicator data changes over the same period.

Since the regression technique measures the growth in SLA population and indicator data relative to the growth in population and indicator data at the state level, the ‘significance’ of change at the SLA level is judged based on the change in the corresponding data item at the state level. SLAs may be analysed further if the modelled population is not supported by indicator data or where the indicator data suggests a greater change than the model estimates.

The overall occupancy ratio for each SLA is analysed as a further check of the balance between population and dwelling numbers. Rapid change in occupancy ratios over time highlights SLAs which may require closer inspection.

Areas known to have experienced major development and/or major events are examined more closely to ensure that such effects have been accounted for by the model, with ad–hoc adjustments made when necessary.

Extra consideration is given to those SLAs which have experienced recent boundary changes in the current intercensal period. Estimates subsequently produced for such areas may be more prone to error than for unaffected SLAs (South Australia underwent a major round of SLA boundary changes in 1998).

In recent times South Australia has been characterised by low levels of population growth. Combined with the absence of SLAs with small populations, or with predominant ‘special’ populations (for example, military or Indigenous populations), most estimates provided by the models do not appear to require significant adjustment.

10.5 Western Australia

A questionnaire is sent annually to all LGAs in Western Australia (WA) requesting advice relating to their population. One of three different questionnaires are provided to each council, based on their basic ‘type’: metropolitan; remote/mining and the remaining country councils. The questionnaire requests a variety of information including electoral population, the number of rateable/non–rateable properties, dwelling activity, economic activity and Indigenous communities. The questionnaire also requests a council estimate of the resident population, and details of how this estimate was arrived at.

The WA Electoral Commission provides, annually, a summary of the number of electors enrolled for each LGA.

The WA modelling process is unique in that three alternative estimates are initially provided, based on the ‘best’ three models (see section 9.5). These modelled estimates are incorporated into a database which also includes the council estimates of usual resident population, the previous year's official population estimate and the number of electors. The database incorporates a time series of these and other (indicator) data. For each SLA, the three alternative modelled estimates are analysed and, utilising local knowledge and the other data, the most appropriate estimate is selected. Graphs and tables are used to assist in the selection process. Each selection, and the reasons for making the particular selection, is documented for consideration in later years.

Once the estimates for all SLAs have been determined it is then necessary, as with all states, to force them to add to the pre–determined state total. A comparison of the selected estimates and pro–rated final figures is made to ensure no final estimate differs considerably to the previous version of the estimate.

10.6 Tasmania

Historical information is kept for all SLAs in Tasmania, relating back to 1993, last phase of widespread SLA boundary changes. Data back to 1986 has been recast to reflect 1993 boundaries and subsequent minor boundary changes, such as those which occurred in July 1996, are progressively applied. Data held for each SLA includes estimated resident population, occupied and unoccupied private dwellings, total dwellings and building approvals. From this, derivations of occupancy ratios, occupied dwelling rates and the rates of conversion of approvals to dwellings are calculated.

Changes in these rates are calculated for each SLA for the current intercensal period. Estimates independent of the regression model are initially calculated and later compared with the modelled estimates. Adjustments may then be made to individual modelled estimates in cases where changes in the population that have occurred over the year have not been adequately reflected.

The ABS, in conjunction with several Tasmanian Government agencies, has formed a Population Working Group for Tasmania. This forum allows for discussion and sharing of relevant information and enables the ABS to obtain details of State Government policy changes which have potential impacts on population.

On an ongoing basis, population intelligence is gathered for the whole of Tasmania. Correspondence with the ABS Population Survey area enables insights to be gained on aspects such as new subdivisions and trends in the number of dwellings for sale. Local media is also a useful source of information on major developments such as call centres.

10.7 Northern Territory

Population intelligence is gathered on an ongoing basis for the whole of the Northern Territory (NT) and, wherever possible, at the SLA level. Sources and types of intelligence gathered for the validation process include:

- the Population Intelligence Working Group (PIWG — made up of representatives of the ABS and various NT Government agencies, who meet quarterly to exchange population intelligence);
- the Department of Local Government (who produces independent service population estimates for various regions);
- the Department of Education (who conduct quarterly surveys on the number of students enrolled in urban schools and an annual census of student enrolments in all NT public schools);
- the media (to gain insight into the opening or closing of mines, opening of or proposals for new residential developments, defence movements, etc)
- the Department of Education, Employment Training and Youth Affairs (who produce the NT Labour Market Profiles);
- the Department of Defence (for number of defence personnel living on the Robertson Barracks and defence movements to and from NT);
- the Defence Corporate Support Centre NT, Kimberley (for personnel numbers and defence movements to and from the NT);
- major Town Councils, especially Palmerston (who independently derive population estimates for SLAs in the Palmerston area) and Alice Springs;
- the Department of Mines and Energy (for mining employment/operations data);
- the NT Statistical Liaison Committee (consisting of ABS and NT Government agencies, which produces population projections by SLA);

- the Real Estate Institute of the NT (for dwelling and rental data);
- the Office of Lands, Planning & Environment (Alice Springs);
- other data (such as ABS mining location employment data correlated to SLA, building approvals, Medicare enrolments, family allowance recipients);
- local knowledge.

The modelled SLA population estimates are analysed based on this intelligence. Each modelled estimate is validated by comparing the most recent year's change in population to the change in previous years. For each SLA, comparisons are made against the indicator data. Adjustments are made to the modelled estimates if they do not appear to adequately reflect changes in population. Comments are sought from PIWG members on potentially suspect SLAs before deciding whether to adjust the modelled estimates.

For SLAs in the Darwin and outer Darwin areas that showed large population changes comparisons were made against information taken from the media, local knowledge, PIWG members, NT Government projections, population estimates by suburb from the Palmerston Town Council, dwelling vacancy rates and rental data. Comparisons were also made against service population estimates, NT Government projections, dwelling vacancy rates and rental data. For mining–based SLAs mining employment comparisons were made against ABS mining employment data, media and the Labour Market Profiles. For other remote SLAs the estimates were validated using Grants Commission service population estimates, media reports, past population trends, building activity data, family allowance data and Medicare data.

10.8 Australian Capital Territory

In the Australian Capital Territory (ACT), SLAs generally equate to suburbs. For validation purposes this is advantageous for some data sources, but negative for others. For example, electricity connections data is available for each suburb, eliminating the need to use geographic concordances to convert data. However in the ACT, where each postcode generally includes a number of SLAs, the quality of indicator data which available by postcode are potentially more adverse to changes in postcode–SLA concordances over time (see section 6.2.1).

SLA boundaries in the ACT have changed little over the past decade, therefore for established suburbs long and continuous time series are available. The ACT is geographically small, which enables more effective local knowledge to be applied.

Data sources used in the production and verification of ACT estimates are: buildings approvals data; domestic electricity connections; Medicare enrolments and ACT Govt Planning and Land Management Group data of quarterly estimates of dwellings (occupied, completed but not occupied, and under construction).

For each SLA, the current estimate is compared with the previous year's estimate and census year estimates to get an indication of change over the past year. For stable SLAs, checking is generally more straightforward. However SLAs with changing populations require more examination of indicator data. Where the indicator data are inconsistent, further investigation is necessary. This includes determining the effect on the estimate of each variable in the regression model, determining SLA characteristics from 1996 Census, discussions with other parties and utilising local knowledge. Changes, if necessary, are made based on an assessment of these factors.

REFERENCES

Australian Bureau of Statistics (ABS). 1977. *A Regression Approach to Small Area Estimation.* Occasional Paper.

Australian Bureau of Statistics (ABS). 1979. *Population Estimates in Australia – a Discussion Paper.* Occasional Paper.

Australian Bureau of Statistics (ABS). 1982. *Regression Techniques for Population Estimation.* Occasional Paper 1982/1.

Australian Bureau of Statistics (ABS). 1983. *Methods and Procedures in the Compilation of Estimated Resident Population 1981 and in the Construction of the 1971–81 Time Series*. Catalogue No. 3103.0.

Australian Bureau of Statistics (ABS). 1985. *Local Government Area Population Estimates, Victoria.* Catalogue No. 3502.2.

Australian Bureau of Statistics (ABS). 1988. *South Australian Intercensal SLA Estimated Resident Population Model.* Unpublished Paper.

Australian Bureau of Statistics (ABS). 1993. *Postcode to Statistical Local Area Concordance, Australia*. Catalogue No. 1253.0.

Australian Bureau of Statistics (ABS). 1996. *Evaluation of Administrative Data Sources for use in Quarterly Estimation of Interstate Migration between 1996 and 2001.* Demography Working Paper 96/1.

Australian Bureau of Statistics (ABS). 1997a. *Census of Population and Housing 1996: Data Quality – Undercount*. Information Paper.

Australian Bureau of Statistics (ABS). 1997b. *Postal Area to Statistical Local Area Concordance, Australia*. Catalogue No. 1253.0.

Australian Bureau of Statistics (ABS). 1998a. *Dwelling Unit Commencements Reported by Approving Authorities, South Australia*. Catalogue No. 8741.4.

Australian Bureau of Statistics (ABS). 1998b. *Dwelling Unit Commencements Reported by Approving Authorities, Western Australia*. Catalogue No. 8741.5.

Australian Bureau of Statistics (ABS). 1998c. *Issues in Estimating Small Area Populations.* Demography Working Paper 98/1 (ABS website http://www.abs.gov.au).

Australian Bureau of Statistics (ABS). 2000. *Demographic Estimates and Projections: Concepts, Sources and Methods.* Statistical Concepts Library, ABS website http://www.abs.gov.au.

Ericksen, E.P. 1974. A Regression Method for Estimating Population Changes of Local Areas. *Journal of the American Statistical Association* 69:867–875.

Goldberg, D., Rao, V.R. and Namboodiri, N.K. 1964. A Test of the Accuracy of the Ratio Correlation Population Estimates. *Land Economics* 40:100–102.

Namboordini, N.K. 1972. On the Ratio–Correlation and Related Methods of Subnational Population Estimation. *Demography* 9:443–453.

Office of Population Censuses and Surveys. 1991. *Making a Population Estimate in England and Wales*. Occasional Paper 37.

O'Hare, W. 1976. Report on a Multiple Regression Method for making Population Estimates. *Demography* 13:369–379.

Purcell, N.J. and Kish L. 1979. Estimations for Small Domains, *Biometrics* 35:365–384.

Rosenburg, H. 1968. Improving Population Estimates with the Use of Dummy Variables*.* *Demography* 7:87–91.

Schmitt, R.C. and Crosetti, A.H. 1954. Accuracy of the Ratio–Correlation Method for Estimating Postcensal Population. *Land Economics* 30:279–281.

Simpson, S., Diamond, I., Tonkin, P., and Tye, R. 1996. Updating Small Area Population Estimates in England and Wales. *Journal of the Royal Statistical Society* 159(2): 235–247.

Snow, E.C. 1911. The Application of the Method of Multiple Correlation to the Estimation of Post–Censal Populations. *Journal of the Royal Statistical Society* 74:575–620.

Statistics Canada. 1992. *Postcensal Annual Estimates of Population for Census Divisions and Census Metropolitan Areas, June 1, 1991 (Regression Method).* Volume 7. Catalogue 91–211.

Statistics Canada. 1995. *Annual Demographic Statistics, 1994*. Catalogue 91–213.

Statistics New Zealand. 1998. Unpublished correspondence.

US Bureau of the Census. 1994. *Evaluation of Postcensal County Estimates for the 1980s.* Technical Working Paper No. 5.

Zitter, M. And Shryock, H.S. 1964. Accuracy of Methods of Preparing Postcensal Population Estimates for States and Local Areas. *Demography* 1:227–241.

APPENDIX 1: Disaggregation of postcensal SLA population totals by age and sex

Following the production of the SLA total populations, attention is focused on the calculation of the age and sex components of the SLA populations. Intercensal estimates of the age and sex distributions of SLA populations are made by updating the population by age and sex for the census year using annual births (by sex) and deaths (by age and sex) data and derived age and sex profiles of migration. While annual data on births and deaths by age/sex are available for each SLA, data on migration for postcensal years are not directly available and have to be derived.

The estimate of total population growth for each SLA for the twelve months is split into natural increase and net migration components. Natural increase is derived for each SLA from birth and death registration data. Net migration is derived for each SLA as the difference between total population growth and natural increase, which can be expressed as:

$$N = P_{t+1} - P_t - B + D \qquad (1)$$

where P denotes the total population, B births, D deaths, and N net migration during the one year from t to t+1.

Net migration is then split into internal and overseas migration components, derived from arrivals and departures data:

$$N = IA - ID + OA - OD \qquad (2)$$

where IA and ID are internal arrivals and departures, and OA and OD are overseas arrivals and departures.

Census year

Migration estimates for each SLA were calculated from the 1996 Census by pro–rating a combination of 1996 Census data on internal movements for 1995–96 and both overseas passenger cards (incoming and outgoing) and 1996 Census data on overseas movements for 1995–96.

The SLA age/sex profiles of internal migration were derived from 1996 Census data based on the SLA of usual residence one year ago. These profiles were produced for inter–SLA arrivals (persons residing in the SLA whose usual residence one year ago was in another SLA) and departures (persons whose SLA of usual residence one year ago was that SLA but whose residence at the date of the census was another SLA).

Each SLA's overseas age/sex arrivals profile was derived from 1996 Census counts for that SLA of people whose usual residence one year ago was overseas. In the absence of age/sex data on overseas departures at the SLA level from either the census or outgoing passenger cards the overseas departure profile for each SLA is assumed to be the same as the overseas arrival profile.

Each migration estimate – obtained by multiplying the total population at time t+1 by a ‘movement rate’ specific for each component and calculated from the 1996 Census movement data for SLAs for the period 1995–96 – can be expressed as:

$$IA = (P_{t+1})\ (IA|) \qquad (3)$$

where IA| is the 'movement rate' for internal arrivals and is obtained from

$$IA| = (IA_{t-1,t})/Pt. \qquad (4)$$

ID, OA, and OD are calculated in the same way.

As census data on usual residence one year ago cannot cover those born less than one year ago, migration for those aged zero was assumed to be half that of one year olds.

Once a first estimate of each of the four migration components was obtained (equation (3)), the inter–SLA age/sex arrival and departure profiles were then constrained so that, for each age and sex, the net effect across all SLAs in a state equalled the final 1995–96 interstate migration estimate. Similarly for overseas arrivals (departures), the total of all SLAs within a state was constrained to the age/sex profile of permanent and long–term arrivals (departures) for 1995–96 for the state. Also, for each SLA, the sum of the four migration components was constrained to the net migration figure (equation (2)). A plus–minus iterative proportion fit (IPF) procedure was used to satisfy these constraints simultaneously (for details on the IPF technique see ABS (2000)).

Non–census years

The census–based SLA age/sex profiles for overseas arrivals and departures and internal arrivals are pro–rated to the SLA overseas arrival and departure and internal arrival totals for the year t to t+1 calculated in the previous step.

However SLA departures are a function of existing SLA population, and so census–based age/sex specific departure rates are employed. As census data, by definition, excludes residents who have left Australia, the requisite data on usual address one year ago is only available for internal (ie inter–SLA) departures, not overseas departures.

SLA age/sex internal departures for the year t to t+1 are obtained by multiplying the population for each single year of age and sex, who survived to the year being estimated, by the age/sex specific departure rate. Thus for each age and sex,

$$ID_{t,\ t+1} = (P_t - D_{t,t+1})\ (ID|) \qquad (5)$$

where P is population and ID| is the internal departure rate and is obtained from the latest census.

ID| was initially calculated using 1996 Census–based data (ie t = 1996) as follows:

$$ID| = (ID_{t-1,t})/(P_t - IA_{t-1,t} + ID_{t-1,t} - OA_{\ t-1,t} + OD_{\ t-1,t}) \qquad (6)$$

The departure rate in (6) is 1995–96 internal departures as a rate of the survived 1995 population (ie the 1996 population with migration removed). This can then be used in (5) to create, say, 1996–97 internal departures by applying it to survived 1996 population. Once an initial age–sex estimate of IA, ID, OA and OD has been obtained for year t+1, an IPF procedure is used to satisfy both the SLA migration component totals (equation (3)) and to ensure that when all SLAs in a state are added, the four components equate to their total state age/sex levels. At the conclusion of each annual iteration of the SLA age/sex estimation process, the SLA–specific overseas and internal arrivals and departure age/sex migration profiles are refined to correct unsustainable migration patterns.

Having established estimates of the migration components, the census–based population estimates for each SLA by age and sex are then updated from the previous year using the component method – that is by adding the births, subtracting the deaths and adding/subtracting the migration (see section 2.2.1) which has occurred in the previous year. The estimates are then validated – with adjustments sometimes made – using a variety of techniques and procedures.

APPENDIX 2: Average absolute intercensal discrepancies, LGAs, 1981 and 1996

In 1981 there were 845 LGAs (excluding NT and ACT). Of these, 34 per cent had a discrepancy greater than five per cent, and 14 per cent had a discrepancy greater than ten per cent. In 1996 these percentages were 23 and 6 respectively.

All Local Government Areas

Year	Number of LGAs	Discrepancy > 5%		Discrepancy > 10%	
		No.	Per cent	No.	Per cent
1981*	845	285	33.7%	116	13.7%
1996	677	156	23.0%	34	5.1%

The decrease in discrepancy from 1981 to 1996 is more apparent when small LGAs (which inherently have high percentage errors) are excluded from the analysis:

LGAs with 1986 population greater than 5,000

Year	Number of LGAs	Discrepancy > 5%		Discrepancy > 10%	
		No.	Per cent	No.	Per cent
1981*	435	100	23.0%	32	7.4%
1996	411	61	14.8%	6	1.5%

* preliminary population estimates for LGAs in the Northern Territory were not made in 1981

APPENDIX 3: Derivation of the 1998 population estimate for Liverpool (C)

Liverpool, with an increase of 4,675 dwellings between 1996 and 1998 was classified within NSW as a high growth urban SLA.

From section 9.1, the appropriate regression equation is:

$$P^{98}_{\text{Liverpool}} = P^{98}_{\text{NSW}} * \left[\frac{P^{96}_{\text{Liverpool}}}{P^{96}_{\text{NSW}}} \right] + 0.47*S_1{}^{96,98}_{\text{Liverpool}} + 0.20*S_2{}^{96,98}_{\text{Liverpool}} + 0.30*S_3{}^{96,98}_{\text{Liverpool}}$$

The data required to estimate the provisional population of Liverpool is:

	Population		*Dwellings*		*Medicare enrolments*			
	1996	1998	Census	Approvals	Persons, aged 0–15		Males, aged 16 +	
Region			1996	1996–98	1996	1998	1996	1998
Liverpool	124 292	(*to be estimated*)	39 371	4 675	31 079	35 117	44 701	50 048
NSW	6 204 728	6 337 876	2 272 244	94 146	1 399 753	1 414 034	2 406 656	2 495 391

Now:

P^{98}_{NSW} = 6,341,594 (preliminary)

P^{96}_{NSW} = 6,204,728

$P^{96}_{\text{Liverpool}}$ = 124,292

$S_1{}^{96,98}_{\text{Liverpool}}$ = change in Liverpool share of NSW **dwellings**, 1996–98

$= (\text{dwellings}^{98}_{\text{Liverpool}} / \text{dwellings}^{98}_{\text{NSW}}) – (\text{dwellings}^{96}_{\text{Liverpool}} / \text{dwellings}^{96}_{\text{NSW}})$

= (44046 / 2366450) – (39371 / 2272294)

= 0.129

$S_2{}^{96,98}_{\text{Liverpool}}$ = change in Liverpool share of NSW **Medicare enrolments, aged 0–15**, 1996–98

$= (\text{M0–15}^{98}_{\text{Liverpool}} / \text{M0–15}^{98}_{\text{NSW}}) – (\text{M0–15}^{96}_{\text{Liverpool}} / \text{M0–15}^{96}_{\text{NSW}})$

= (35117 / 1414034) – (31079 / 1399753)

= 0.263

$S_3{}^{96,98}_{\text{Liverpool}}$ = change in Liverpool share of NSW **Medicare enrolments, males aged 16+**, 1996–98

$= (\text{Mm16+}^{98}_{\text{Liverpool}} / \text{Mm16+}^{98}_{\text{NSW}}) – (\text{Mm16+}^{96}_{\text{Liverpool}} / \text{Mm16+}^{96}_{\text{NSW}})$

= (50048 / 2495391) – (44701 / 2406656)

= 0.015

= 137,047

This figure is then scrutinised by the local ABS office, which validates this figure and its associated change over time together with a range of other indicator data. No particular adjustment was made to Liverpool in 1998.

All other SLAs are checked this way. The combined effect of the adjustments made for all other SLAs in NSW (ie given that the sum of the SLA populations needs to equal the pre–determined state population) was a small increase (19 persons) for Liverpool.

The preliminary 1998 ERP for Liverpool (C) was 137,066.

APPENDIX 4: Average absolute discrepancy using regression models to estimate populations of SLAs, with and without stratification, 30 June 1996

	Average absolute discrepancy			
	Percentage (per cent)		Numeric (persons)	
	No stratification	Stratification	No stratification	Stratification
New South Wales	2.28	1.97	510	464
Victoria	3.04	2.89	497	415
Queensland	5.45	5.37	286	277
South Australia	3.03	2.94	169	161
Western Australia	6.85	6.03	335	270
Tasmania	3.33	2.66	201	158
Northern Territory	8.47	7.85	159	136
Australian Capital Territory	4.48	3.79	91	70

APPENDIX 5: Strata developed for the 1997–2001 SLA estimation models

State	**Location**	**Low growth (increase in dwellings)***	**High growth (increase in dwellings)***	**Urban / rural definition** *SLAs within the following regions**:*
NSW	Urban	< 1100	≥ 1100	Sydney SD, Lake Macquarie, Newcastle (Inner and Remainder), Shoalhaven & Wollongong SLAs
	Rural	< 580	≥ 580	Rest of state
Vic.	Urban	< 1150	≥ 1150	Melbourne SD, Greater Geelong City Pt A SSD
	Rural	< 600	≥ 600	Rest of state
Qld	Urban	< 47%	≥ 47%	Brisbane SD, Moreton SD
	Rural	< 25%	≥ 25%	Rest of state
SA	Urban	< 1100	≥ 1100	Adelaide SD
	Rural	< 200	≥ 200	Rest of state
WA	Urban	< 1500	≥ 1500	Perth SD
	Rural	< 500	≥ 500	Rest of state
Tas.	Urban	< 14%	≥ 14%	Greater Hobart, Greater Launceston, Greater Burnie–Devonport SSDs
	Rural	< 1.9%	≥ 1.9%	Rest of state
NT	Urban	< 40	≥ 40	Darwin SD, Alice Springs LGA
	Rural	< 40	≥ 40	Rest of state
	Other			
ACT***	High flats	< 40	≥ 40	'High flats' SLAs are those where ≥18% of the dwellings in the SLA are not houses
	Low flats	< 60	≥ 60	'Low flats' SLAs are those where <18% of the dwellings in the SLA are not houses

* Growth thresholds were determined based on the increase in dwellings between 1991 and 1996, where data from the 1991 Census of Population and Housing was used as the base, and was updated to 1996 with building approvals for all states except for ACT, who used electricity connections.

** SD denotes Statistical Division; SSD denotes Statistical Subdivision.

*** ACT was not split into rural and urban due to the overwhelmingly urban nature of the state.