Explanatory Notes

2015 modelled estimates for small areas based on 2015 Survey of Disability, Ageing and Carers

Prepared by Small Population Insights, Australian Bureau of Statistics

Release: May 2020

LICENSE CONDITIONS

This customised report carries the following licence:
Creative Commons Attribution 2.5 Australia

You are free to re-use, build upon, and distribute this material, even commercially. The entire report may be included as an appendix in your work for reference if you wish.

Under the terms of this license, you are required to attribute ABS material in the manner specified (but not in any way that suggests that the ABS endorses you or your use of the work).

ABS material used 'as supplied'
Provided you have not modified or transformed ABS material in any way, for example by:

- changing the ABS text
- calculating percentage changes
- graphing or charting data
- deriving new statistics from unpublished ABS statistics

material contained in this customised report may be reused provided one of the following attributions is given:

Source: Australian Bureau of Statistics
or
Source: ABS

Derivative material
If you have modified or transformed ABS material, or derived new material from those of the ABS in any way, one of the following attributions must be used:

Based on Australian Bureau of Statistics data
or
Based on ABS data

Citing customised reports
If you are required to cite material from this report please be guided by the examples below.

In-text and reference list/bibliography

- In-text:
  (ABS 2020)
- In reference list/bibliography:
  ABS 2020, Customised report.

In-text only

- (ABS, Customised report, 2020)

CONTENTS

1. INTRODUCTION

The Australian Bureau of Statistics (ABS) has produced a core set of modelled estimates of characteristics associated with disability at a small area level for the Australian population, based on the Survey of Disability, Ageing and Carers (SDAC). These explanatory notes accompany the modelled estimates for small areas, provided as Excel worksheets, and describe the methodology used to produce them, as well as how to use them.

2. PURPOSE

The SDAC, a national sample survey, collected a range of information from Australians about disability, need for support for older people and those with a disability, and information about carers, to enable reliable estimates at the national, state and territory level. The sample size was too small to produce reliable estimates for areas with small populations such as Statistical Areas Level 2 (SA2s) or Local Government Areas (LGAs). To produce reliable and detailed estimates at these geographical levels, models were created using the detailed SDAC data and applied to data for small areas on ABS Estimated Resident Population (ERP), the 2016 ABS Census of Population and Housing, and administrative sources.

A modelled estimate can be interpreted as the likely value for an area based on the demographic information we have for that area.

3. METHODOLOGY USED

The process of producing modelled estimates at the SA2 and LGA level on indicators measured in the SDAC consisted of the following components:

1. Identification of the outcome variables
2. Selection of the predictor variables
3. Identification of the geographical regions
4. Identification of population estimates
5. Scoping the data
6. Creation of binary and proportion variables
7. Aggregating observations and merging datasets
8. Model selection
9. Creation of modelled estimates
10. Assessment of the modelled estimates

3.1 Identification of the outcome variables

The 2015 SDAC is the most recent in a series of comprehensive national surveys conducted by the Australian Bureau of Statistics, following similar surveys in 1981, 1988, 1993, 1998, 2003, 2009 and 2012. The survey was conducted in all states and territories and across urban, rural and remote areas of Australia (other than very remote areas). Data collection was in two parts: the establishment component, which was conducted from 25 May to 31 July 2015; and the household component, which ran from 5 July 2015 to 19 December 2015. Accommodation within establishments included hospitals, nursing homes, aged care hostels, cared components of retirement villages and other 'homes', such as group homes for people with disability. The survey included around 63,500 people from over 25,500 private dwellings, and a further 11,700 people from 1,000 establishments.

The survey was designed to:

- measure the prevalence of disability in Australia
- measure the need for support of older people (those aged 65 years and over) and those with disability
- provide a demographic and socio-economic profile of people with disability, older people and carers compared with the general population
- estimate the number of, and provide information about, those who provide care to people with disability and older people.

Indicators for modelling are often referred to in literature as outcome variables, dependent variables or response variables. From the 2015 SDAC, modelled estimates (counts, proportions, error) about people from the overall Australian population (excluding establishments component of the SDAC and very remote areas) with the following characteristics were produced.

At the SA2 and LGA level:

Degree of restriction by age and sex
- Any disability
- Profound/Severe
- Moderate/Mild

Carer status by age and sex
- All carers
- Primary carer
- Non-primary carer

For age groups:
- 0-14
- 15-24
- 25-34
- 35-44
- 45-54
- 55-64

- 65+

For more information about the outcome variables, including definitions, see the worksheet 'Notes' within the MS Excel spreadsheets.

More information about the 2015 SDAC can be found on the ABS website, www.abs.gov.au; for example, Disability, Ageing and Carers, Australia: Summary of Findings, 2015 (ABS catalogue number (4430.0) includes information about the collection of the survey and results.

3.2 Selection of the predictor variables

In order to predict outcome variables, predictor variables are required on both the SDAC dataset and a small area dataset containing population, Census, and administrative data. Predictor variables are also referred to in literature as explanatory variables or independent variables. Predictor variables were created if data were available for small areas for all of urban, rural, and remote Australia and if there was an expectation that they might be good predictors of the outcome variables.

For age and sex predictor variables, data at the small area level were obtained from ABS ERP data from ABS Regional Population Growth, Australia, 2015-16 (Cat. No. 3218.0) (described below in section 3.4).

For other demographic variables on the SDAC, data at the small area level were obtained from the 2016 Census of Population and Housing, as this was the most up-to-date comprehensive source of demographic data due to the depth of information at small geographical levels.

If appropriate, demographic variables on the SDAC and Census were adjusted to make them more closely align. Variables that were available at the small area but not collected in the SDAC were added to the SDAC data; these variables included other demographic variables on the Census, geographic variables, and variables from administrative sources.

Predictor variables that relate to the geographical areas where people reside included:

- remoteness area;
- socio-economic indexes for areas (SEIFAs) – population-weighted deciles at the Statistical Areas Level 1 (SA1) level;
- state and territory;
- section of state (major urban/other urban/bounded locality/rural balance);
- Greater Capital City Statistical Area (GCCSA)/balance of state; and
- design area type (categorises inner city, large and small urban towns, rural towns and remote areas within states and territories for designing the sample of the SDAC).

Sources of data included:

- ABS Australian Statistical Geography Standard (ASGS): Volume 5 - Remoteness Structure, July 2016 (Cat. no. 1270.0.55.005);
- Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2016 (Cat. no. 2033.0.55.001);
- Australian Statistical Geography Standard (ASGS): Volume 4 - Significant Urban Areas, Urban Centres and Localities, Section of State, July 2016 (Cat. no. 1270.0.55.004); and
- Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2016 (Cat. no. 1270.0.55.001).

Predictor variables from administrative sources included:

- births in 2015;
- deaths in 2015;
- dwelling transfers and median sale prices in 2014-15;
- immigration from other areas of Australia in 2014-15;
- income inequality measures for total income earners 2014-15;
- attendance at public hospitals for various conditions and procedures (2014-15), deaths from selected causes (2010 to 2014), Home and Community Care Program (HACC) clients (2014-15), participation in vocational education and training (2015), development of children (2015), immunisations (2015), and bowel cancer screening (2014-15).

Data for the above were obtained, in the same numeric order, from:

- ABS Births, Australia, 2016 (Catalogue number 3301.0)
- ABS Deaths, Australia, 2016, (Catalogue number 3302.0)
- ABS Data by Region, 2012-17 (Catalogue number 1410.0)
- ABS Migration, Australia 2015-16 (Catalogue number 3412.0)
- ABS Estimates of Personal Income for Small Areas, 2011-16 (Catalogue number 6524.0.55.002)
- Public Health Information Development Unit (PHIDU) http://phidu.torrens.edu.au, March 2018 release.

Within most types of predictor variables (as discussed above), several explanatory variables representing separate categories or data items were tested. The variables tested for significance are listed in Appendix 1 and in the MS Excel spreadsheets.

3.3 Identification of the geographical regions

The geographical regions required were at a level where direct survey estimates are not available or their sample errors are too high for the direct survey estimates to be useful. Modelled estimates were provided at the SA2 and LGA level.

Statistical Areas Level 2 (SA2s) are defined within the main structure of the Australian Statistical Geography Standard (ASGS). They are medium-sized general purpose areas built up from whole Statistical Areas Level 1. Their purpose is to represent a community that interacts together socially and economically.

There are 2,310 SA2 regions covering the whole of Australia without gaps or overlaps. These include 18 non-spatial SA2 special purpose codes, comprising Migratory–Offshore–Shipping and No Usual Address codes for each State and Territory. See Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, July 2016 (Catalogue number 1270.0.55.001) for more information about the ASGS.

Local Government Areas (LGAs) are not defined or maintained within the main structure of the ASGS. They are an ABS approximation of gazetted local government boundaries as defined by each State and Territory Local Government Department. Local Government Areas cover incorporated areas of Australia. Incorporated areas are legally designated parts of a State or Territory over which incorporated local governing bodies have responsibility. The major areas of Australia not administered by incorporated bodies are the northern parts of South Australia, and all of the Australian Capital Territory and the Other Territories. These regions are identified as 'Unincorporated' in the ASGS Local Government Areas structure. See Australian Statistical Geography Standard (ASGS): Volume 3 – Non ABS Structures, July 2018 (Catalogue number 1270.0.55.003) for more information about the ASGS.

The 2018 ASGS edition of LGAs has been used.

3.4 Identification of population estimates

ABS ERP data from ABS Regional Population Growth, Australia, 2015-16 (Cat. No. 3218.0) were used.

This data was then adjusted to match the scope of the SDAC and to sum to population SDAC state by age by sex estimates (described below in section 3.5).

The adjusted ERP data were also used as denominators in the calculations of proportions of persons i.e. 'population' estimates included in the MS Excel spreadsheets. It is important to note that these population estimates are not official estimates and were created solely for analysis of the SDAC modelled estimates and will not match other population data for SA2s or LGAs.

3.5 Scoping the data

The modelled estimates for small areas are applicable to persons who were usual residents living in households to match the scope of the SDAC. They exclude areas that have been classified as very remote or as Indigenous communities.

Adjustments were made to the ERP data, by using ratios of persons living in households to persons not living in households, calculated from the 2016 Census to approximate exclusion of persons not living in households, and then summed to the SDAC population state by age by sex benchmarks. These are the 'population' estimates included in the MS Excel spreadsheets.

Adjustments were also made to the Census data, specifically the predictor variables obtained from the Census (described above in section 3.2). Identification of a persons' type of dwelling is possible on the Census datasets for respondents at home on Census night so persons not living in households were easily removed from the small area dataset. However, for persons who were not at home on Census night, information is not collected to determine if the dwelling they usually reside in is a household or not; therefore, their records were deleted from the small area dataset. This means that an assumption has been made that the people who were away from home on Census night and live in households have the same disability characteristics as the people who were at home in households.

Removal of very remote and Indigenous communities from the ERP and Census data file was approximately done by deleting persons residing in SA2s/LGAs that had more than 20% of their population in SA1s classified as very remote or in Indigenous communities. See the worksheets "Excluded SA2s" and "Excluded LGAs" within the MS Excel spreadsheets for the SA2s/LGAs excluded.

Other territories and SA2s/LGAs with no population were also excluded.

Foreign diplomatic personnel and their families were excluded from the modelled estimates because they are not included in Australia's ERP, the Census or the SDAC.

While out of scope for the SDAC, members of non-Australian defence forces (and their dependents) stationed in Australia were unable to be removed from the modelled estimates because they could not be identified in Australia's ERP.

For more information on the survey scope and coverage, see the Disability, Ageing and Carers, Australia: First Results, 2015 (ABS catalogue number 4430.0.10.001).

3.6 Creation of binary and proportion variables

On the SDAC dataset outcome variables were created as binary variables to make them suitable for the type of modelling undertaken (logistic regression). On both the SDAC and small area datasets, predictor variables that were categorical were also created as binary variables. An observation took the value of 1 if an individual had a characteristic of interest and 0 if otherwise. For example:

1. in the case of persons with a disability with a profound or severe core activity limitation, the outcome variable for this took the value of 1 if an individual had a profound or severe disability and 0 if the individual did not have a profound or severe disability; and

2. in the case of labour force status, the predictor variable for employed took the value of 1 if an individual was employed and 0 if the individual was unemployed, not in the labour force or aged 0-14 years.

Variables in administrative data were converted to proportions of their areas' population and attached to the SDAC and small area datasets. For example, a person can live in an area with a proportion of its population receiving a disability support pension (a value of 1) or not (a value of 0). Some categorical variables (e.g. ranges) are also derived and then converted into binary variables, as above, before being attached. For example, a person can live in an area with a fertility rate between 2 and 2.5 (a value of 1) or not (a value of 0).

3.7 Aggregating observations and merging datasets

All the datasets were aggregated to combine like observations/respondents based on areas of usual residence (e.g. SA2, LGA), design area type, five year age groups and sex. This decreases the size of the datasets (especially the Census dataset) to increase the efficiency of the modelling process.

The Census, adjusted ERP and administrative datasets were merged into one small area dataset. A number of the areas of usual residence by age by sex groups had a non-zero adjusted ERP with no corresponding combination within the Census dataset. For the most part this was due to population growth between the time the Census was undertaken and when the adjusted ERP was created, or because areas had very small populations; for all but a small number of the affected areas the effect is insignificant. Given that Census data is required in order to derive appropriate modelled estimates, the affected groups have been excluded from estimates at SA2 and LGA level. It is not expected that these exclusions will have a significant impact on modelled estimates at the SA2 and LGA level.

3.8 Model selection

Models were created for each outcome variable independently. The model selection method uses the data files to measure the relationship between the outcome variable and possible predictor variables to determine one set of significant predictor variables. This method assumes that the relationships observed in the survey data overall also hold at the small area level.

The models used to determine these relationships were logistic regression models. As part of any model selection process an appropriate significance level must be chosen for determining which explanatory variables to include in the models. The 0.05 (95%) level is most commonly used; however, due to SDAC's relatively large sample size, the Bayesian Information Criterion (BIC) was used to reduce the risk of over-fitting.

The models were applied to small area data using 2015 adjusted ERP for the population counts (described above in section 3.4), summed to create Australia level modelled estimates and compared with reliable direct survey weighted estimates to see if the model adequately predicted the outcome variable. Some models were improved with the addition of less significant predictor variables and interactions of some predictor variables.

3.9 Creation of modelled estimates

The relationships selected above were then fitted using random effects logistic regression models. A mixed estimate comprised of modelled and survey data is then produced for each SA2 and LGA. A mixed/composite estimate gives results for each small area that reflect the best trade-off between the accuracy of the direct survey weighted estimate and the error associated with the modelled estimate. So, for a small area that happens to have a low sampling error (because of a large sample size within that small area, for example), more weight will be given to the direct estimate when calculating an estimate for that small area. On the other hand, for a small area with high sampling error, more weight will be given to the model based prediction as this will be more reliable in calculating the estimate for that small area. This takes advantage of what is known about SA2s and LGAs from the survey to improve the modelled estimates.

A pro rata adjustment was then made to the modelled estimates so that they summed to state and territory direct survey estimates and broken down by age or by sex if applicable. The associated errors resulting from the modelling process, which improve on direct survey estimates' errors, were not adjusted.

The SA2 and LGA modelled estimates for each age and sex group had different adjustments; therefore, the sum of these modelled estimates by age or by sex will not exactly sum to their totals.

The modelled estimates supplied in the MS Excel spreadsheets are in the form of counts (number of persons) and their relative error for the SA2s/LGAs. Prevalence proportions (percentage of population in each small area) and their confidence intervals have been calculated. The denominators used in the calculation of proportions were the unofficial population estimates for each SA2/LGA (based on adjusted ERP) described above in section 3.4.

3.10 Assessment of the modelled estimates

Various measures were taken to examine the modelled estimates. Modelled estimates were compared with direct survey estimates from the 2015 SDAC for areas that were sampled. For the survey estimates, 95% confidence intervals were calculated. These were plotted against the modelled estimates to see if the majority of modelled rates fell within the confidence intervals of the SDAC estimates. Also, relative root mean squared errors (RRMSEs) of the modelled estimates were examined to ensure that the majority were of reasonable quality.

4. ACCURACY OF RESULTS

The process undertaken in providing modelled estimates overcomes much of the volatility at the SA2/LGA level caused by sampling error. However, it should be remembered that the estimates provided are still subject to errors.

The errors associated with the modelled estimates for small areas fall into four categories, as follows.

4.1 Sampling Error

Sampling error is introduced into estimates because SDAC data were collected for only a sample of dwellings. Therefore, they are subject to sampling variability; that is, modelled estimates may differ from those that would have been produced if all dwellings had been included in SDAC. Furthermore, the smaller the sample obtained within a small area, the greater the sampling error associated with that small area's modelled estimates will be.

4.2 Non-Sampling Error

The imprecision due to sampling error should not be confused with inaccuracies that may occur because of imperfections in reporting by respondents and recording by interviewers, and errors made in coding and processing data. Inaccuracies of this kind are referred to as non-sampling error, and they occur in any enumeration, whether it be a full count (census) or a sample. Unlike the other sources of error, non-sampling error is not measurable and therefore isn't accounted for in the measured error (direct or modelled) that accompanies ABS estimates. Every effort is made to reduce non-sampling error to a minimum by careful design of questionnaires, intensive training and supervision of interviewers, and efficient procedures.

4.3 Modelling Error

Modelling error is introduced by model misspecification. This can occur when the choice of model is incorrect, a key explanatory variable is left out or an inappropriate explanatory variable is included. In practice, it is rarely the case that all determinants of health indicators will be available as good quality small area data to be able to be included as predictor variables in the models. Therefore, the variables chosen in the models may result in incorrect modelled estimates for certain small areas, particularly those unusual small areas that do not follow the typical associations between the available predictor variables and the health indicators. The models that have been chosen have been tested against a range of possible alternative models; however, they are only the most preferred models subject to available predictor variables.

4.4 Prediction Error

A strong model does not guarantee statistically accurate modelled estimates. Prediction error is a measure of the statistical accuracy of the model predictions. A measure of the quality of the modelled estimates is the relative root mean squared error (RRMSE). The RRMSE is primarily a measure of prediction error but in its calculation it also inherits some aspects of modelling and sampling error. The RRMSE generally decreases as the population size increases, and is used to assess the reliability of modelled estimates.

As a general rule of thumb, estimates with RRMSEs less than 25% are considered reliable for most purposes, estimates with RRMSEs between 25% and 50% should be used with caution and estimates with RRMSEs greater than 50% are considered too unreliable for general use.

Some areas/groups have high RRMSEs and the accuracy of their modelled estimates can be improved by aggregating them to larger regions/groups. The method and examples for calculating RRMSEs for aggregated areas/groups are provided in the MS Excel spreadsheets.

A confidence interval (CI) provides a range of values, within which it is estimated that the true population value lies. To assist with the calculation of confidence intervals for aggregated areas, examples of this calculation are provided in the Excel spreadsheets.

5. USING MODELLED ESTIMATES

The small area modelled estimates can be interpreted as the expected prevalence for a typical area in Australia with the same characteristics. For some SA2s and LGAs there will be large differences between the modelled estimates and the actual number of people with the characteristic of interest. One explanation for this is that significant local information about particular SA2s and LGAs exists but has not been collected for all areas and cannot be incorporated into the models. This sort of information is usually not measurable, and relies on local or expert knowledge.

Small area modelled estimates should be viewed as a tool that when used in conjunction with local area knowledge as well as the consideration of the modelled estimates reliability, can provide useful information that can assist with decision making for small geographic areas.

Estimates have been confidentialised to ensure they meet ABS requirements for confidentiality. SA2s/LGAs with populations or modelled counts that didn't meet the confidentiality rules have modelled estimates comprised solely of the modelled component; no sampled contribution is included, regardless of whether sample exists in these SA2s/LGAs. The remaining SA2s/LGAs have modelled estimates that are a mix of modelled and sampled components as discussed in 3.9 above.

Areas or groups can be aggregated together using examples provided in the Excel spreadsheets. Example 1 provided shows how to aggregate RRMSEs of several small areas. Example 2 provided shows how to aggregate RRMSEs within a small area (e.g. age groups). Aggregation of small areas should be done taking into account local knowledge about these areas.

The reliability of the resulting aggregated estimate should be assessed in terms of the error values, confidence intervals, and what is known about the 'new' small area or aggregation group.

APPENDIX 1 – LIST OF PREDICTOR VARIABLES CONSIDERED

This summarises the types of variables that were tested for significance in the models.

Relating to persons:

- sex
- age
- elapsed years since arrival in Australia
- country of birth
- country of birth of father
- country of birth of mother
- main field of highest non-school qualification
- Indigenous status
- industry of employment
- labour force status
- occupation
- level of highest non-school qualification
- registered marital status
- relationship in household
- highest year of school completed
- attends University or TAFE
- hours usually worked each week
- personal income
- ancestry
- ancestry
- employee or owner
- employment sector
- field of qualification
- needs assistance for core needs
- number of children ever born to female
- number of employees of owner businesses
- provided unpaid assistance to a person with a disability
- religion
- social marital status
- unpaid domestic work
- volunteer

Relating to the dwelling that persons reside in:

- number of bedrooms
- dwelling structure
- landlord type
- tenure type
- household composition/type
- equivalised gross weekly household income
- household income
- household with Indigenous persons
- number of persons usually resident
- number of children in household
- family blending
- family type
- household five year mobility indicator

- household has an Internet connection
- household one year mobility indicator
- mortgage amount
- rent amount
- whether at the same address five years ago
- whether at the same address one year ago
- labour force status of family

Relating to areas where persons reside in:

- remoteness area
- state and territory
- greater capital city statistical area (GCCSA)
- design area type (a)
- SEIFA Index of Economic Resources (IER) (b)
- SEIFA Index of Education and Occupation (IEO) (b)
- SEIFA Index of Relative Socio-Economic Disadvantage (IRSAD) (b)
- SEIFA Index of Relative Socio-Economic Disadvantage (IRSD) (b)
- Numbers of births and deaths, and fertility rates
- dwelling, house and unit sales
- median house, unit and dwelling sales prices
- income inequality measures for total income earners
- population density (persons/sq km)
- arrivals from internal migration
- participation in vocational education and training
- children developmentally at risk or on track in selected domains
- children fully immunised
- HPV vaccine coverage
- participation in the National Bowel Cancer Screening Program (NBCSP)
- hospital admissions for selected causes
- average avoidable deaths from selected causes.

(a) Design area type categorises inner city, large and small urban towns, rural towns and remote areas within States and Territories for designing the sample of the SDAC.
(b) Socio-economic indexes for areas (SEIFAs) – population-weighted deciles at the Statistical Areas Level 1 level (SA1).

APPENDIX 2 – QUALITY SUMMARY FOR MODELLED ESTIMATES

Measures of prediction accuracy (RRMSEs and CIs) are included in the output provided and can be used to assess the overall reliability for each of the models. The average RRMSE across small areas was calculated for each of the six models used to construct the three tables of output.

When determining the overall reliability of each model, average RRMSEs less than 25% have been summarised as 'reliable'. Average RRMSEs between 25% and 50% have been summarised as 'use with caution'. Models with average RRMSEs greater than 50% have been summarised as 'unreliable for general use'.

Most modelled estimates have been flagged as 'use with caution', based on the average RRMSEs across SA2s/LGAs.

Table	**Average RRMSE (SA2)**	**Average RRMSE (LGA)**	**Reliability**
Disability indicators:			
Table 1 Disability, by sex by age	25.6	19.9	Use with caution at SA2 level
Table 2 Profound or severe core activity limitation, by sex by age	34.3	28.8	Use with caution
Table 3 Moderate or mild core activity limitation, by sex by age	30.5	24.7	Use with caution at SA2 level
Carer indicators:			
Table 4 All carers, by sex by age	34.3	27.2	Use with caution
Table 5 Primary carers, by sex by age	28.7	24.8	Use with caution at SA2 level
Table 6 Non-primary carers, by sex by age	38.6	30.6	Use with caution

The distribution of the estimates across the areas within age groups and genders was as expected for all outcome variables. For example, modelled estimates for disability were consistently higher for the older age groups than the youngest.

It is also useful to consider the number, range, and applicability of explanatory variables included in the models used to create the small area estimates. All of the outcome variables had a good range of explanatory variables included in the models.