Notes on method: the BITRE's taxable income database 2007 update

This is a interactive pdf,
please click on buttons
on the bottom of pages.

Introduction

The Bureau of Infrastructure, Transport and Regional Economics (BITRE) maintains a database of taxable income indicators, classified by region using the Australian Standard Geographical Classification (ASGC) (ABS 2001). Data are derived from Australian Taxation Office (ATO) individual income data, published on a postcode basis in the ATO's annual Taxation Statistics publications.

The BITRE first released its taxable income database in 2005 for the years 1980–81 to 2000–01. Notes on the method used for that period can be found in the accompanying information paper, Focus on Regions 3: Taxable Income, at http://www.bitre.gov.au/publications/22/Files/IP54.pdf.

The 2004–05 update includes revised estimates for the indicators in the database from 2001–02 onwards. These revisions have been made in the light of the additional information published by the ATO in Taxation Statistics 2004–05 (2007). The revisions generally occur in smaller statistical local areas (SLAs).

The ATO's Taxation Statistics series includes taxable income estimates classified to unknown postcodes (the 'other' category). The first stage of the BITRE estimation process involved identifying postcodes used elsewhere in the time series, imputing values for unknown years for these postcodes and adjusting the 'other' category estimates accordingly. The second stage involved transforming the geographic classification used from postcode to ASGC 2001 at the SLA level.

The updated data were estimated using a two step concordance process. This differs from the methodology for years up to and including 2000–01 which were estimated using a single step concordance. The two step process is less satisfactory because it reduces differences between regions. The BITRE and the Australian Bureau of Statistics (ABS) are developing single step concordances for the later years. These will be available for use in next year's taxable income database update.

All monetary values have been adjusted for inflation using the Consumer Price Index (ABS cat. no. 6401.0), as published by the Reserve Bank of Australia (2007), and are presented in 2006–07 equivalent values (2006–07 dollars).

Missing values

The postcode data published by the ATO for the four years 2001–02 to 2004–05 contain a catch-all 'other' category for each state and territory (the 'unknowns'). This category contains data ascribed to invalid or unknown postcodes as well as data from postcodes with small numbers of taxpayers (less than 50 in 2004–05 and 2003–04, less than 5 prior to that back to 2000–01) omitted by the ATO to ensure the confidentiality of individual taxpayers. In addition, the body of the ATO tables present data for 'residential' postcodes but excludes some postcodes specifically assigned to post office boxes.

In the interests of maintaining data continuity across years, it is necessary to make estimates for postcodes where data are missing. To not do so would generate errors where:

- postcodes moved above or below the minimum threshold for publication set by the ATO.
- post office boxes have been transferred from a 'residential' postcode to one specifically reserved for post office boxes.

Identifying postcodes with missing values

Estimates were only compiled for postcodes where there was at least one value in the period from 2000–01 to 2004–05 and some evidence that the postcode was still in use in at least one other year – either in the Australia Post list of postcodes or in the ATO list of concordances.

There were 416 postcodes with between one and four missing values over the five year period.

Number of postcodes with missing values over the 5 years

Number of missing values for the postcode	*Count*
1 value missing	96
2 values missing	140
3 values missing	57
4 values missing	123
Total	416

Average size of postcodes with missing values (number of taxable individuals)

Number of taxable individuals (averaged over the 5 years)	*Count*
0–49	47
50–99	225
100–299	94
300–999	40
1,000–2,452	10
Total	416

The initial missing values were identified as either:

- 'missing', perhaps due to falling below the ATO threshold for publication, not being regarded as 'residential' or simply an error. These postcodes were classed as warranting estimation.
- not genuinely missing, due to the creation or closing down of the postcode. This was determined by the absence of the postcode from any of the ATO lists, the Australia Post list of postcodes and the ABS's list of concordances for that year. Postcodes not appearing on any of these lists were assumed to have not commenced or been phased out and so not estimated. Some 14 postcodes that only had a single value in the ATO data (all in the year 2002–03), had not been estimated in 2000–01 and could not be found in any of the other sources. Values for these postcodes (4056, 4057, 4071, 4126, 5058, 3157, 6115, 2613, 5078, 6724, 4203, 4206, 3291 & 4523) were not estimated for the missing years. The 2002–03 values for these postcodes were added back to the 'unknowns' in their respective states.

In addition to 'missing' postcodes, two postcodes that had values (2890 & 6958) were identified as relating to the Australian Defence Forces, Sydney and Royal Australian Navy Warships respectively. The values from these postcodes were **added to** the 'unknown Australia' category in the years that they were present.

Estimating missing values

Missing values between years of known values:

Where postcodes were missing data between years with known values, estimates were made for intervening year(s) on the basis of a straight line trend between the known values. Note that this simple straight line trend was used across all parameters: no allowance was made for inflation, etc.

If the evidence trail (from any of the ATO, Aust Post & ABS sources cited above) suggested that a postcode had not yet been created or had been abolished, taxable income was set to zero for the years the postcode was assumed to not exist.

Note that the 2000–01 published values were retained and treated as real values for this purpose, even if they were known to have been calculated. Informal sensitivity testing of the effects of recalculating 2000–01 and earlier values in the light of later data revealed little change in the resulting values. For simplicity, the 2001 values have been allowed to stand.

Data missing from the end of the series:

In many cases there were missing data even though there was evidence that the postcode was still in use in 2004–05 (and perhaps beyond) – i.e. the missing part of the series was unbounded. This missing data could not be calculated using the linear trend technique described above. In these cases, the data were assumed to move in the same direction and at the same rate as the data in a 'similar' postcode. The 'similar' postcodes to be used for this calculation were identified on the basis of their location (usually adjacent to the one requiring estimation) and known characteristics (ie rural/urban etc). Post office box based postcodes were linked to the postcode in which they were physically located.

Comparison of estimated "unknown" (residual) postcode values

The original ATO data in each state and territory contain data where the postcode is unknown ('other'). As well as genuinely unknown postcodes, this category contains those postcodes not published due to privacy concerns – that is, those falling below the ATO publication threshold. By estimating these postcodes, the BITRE has reduced the size of the 'unknowns' in each state.

The following table sets out the remaining number of taxable individuals (NTI) in the 'unknown/other' category for each state and territory after BITRE estimation, compared to the number shown in the ATO data for each year.

The differences between these totals can also be affected by the transformations from postcode to SLA geography. Net gains or losses to the totals can occur since, in the ATO data, postcodes crossing state borders are assigned to one state only. The concordance to SLA, however, divides the postcodes which cross borders among the relevant states.

Note that whilst all the values shown in the following table are positive for NTI, the BITRE estimates of the 'unknowns' for the 'non-taxables' indicator in South Australia contain negative values.

Number of taxable individuals in the 'unknown/other' category, ATO and BITRE estimates

	2001–02 NTI		*2002–03 NTI*		*2003–04 NTI*		*2004–05 NTI*	
	ATO	*BITRE*	*ATO*	*BITRE*	*ATO*	*BITRE*	*ATO*	*BITRE*
NSW	6 561	4 253	7 069	3 518	12 850	3 177	13 420	3 144
VIC	5 012	3 313	4 876	3 031	10 465	2 796	10 395	2 497
QLD	3 521	2 149	2 945	1 982	5 305	1 814	5 150	1 040
SA	2 495	143	2 604	126	8 580	229	2 175	173
WA	6 133	2 585	5 392	2 110	11 655	1 839	11 920	1 381
TAS	940	602	974	516	980	444	1 030	496
NT	237	1 291	207	1 372	10 535	1 318	12 340	727
ACT	308	214	259	224	960	212	905	157
Australia*	30 867	20 279	29 686	18 283	66 553	17 052	62 095	14 375

* Includes overseas and state/territory unknowns.

Concordances

The data have been transformed for each year from postcode geography to statistical local area (SLA) geography, using ABS concordances.

This transformation was a two step process for each year. First the data was concorded from postcode geography to the SLAs that existed in the year of collection, and then concorded again to the 2001 ASGC SLA boundaries. The finished dataset therefore has standard ASGC 2001 geography across years.

A number of the postcodes with ATO data were not in the ABS-supplied postcode to SLA concordances for each year. It was therefore necessary for the BITRE to estimate some additional concordances for these postcodes.

Post boxes

For postcodes assigned to post office boxes, the BITRE created concordances identical to the concordances for the areas within which the post boxes were located.

Delivery areas and Large Volume Receivers (LVRs)

For delivery areas, an ABS concordance was identified from the nearest possible year to the one in question. In the very small number of cases where such an ABS concordance could not be identified, the postcode was allocated completely to the SLA containing the main areas of settlement. Happily, in all cases there was no need to divide these postcodes (ie all settlements within the postcode fell within one SLA).

For LVRs (large volume receivers), the location of the postcode collection locality (as identified from Australia Post data) was used. This data was concorded on the same basis as the physical postcode in which it was located.

Data quality

Preliminary estimates were assessed to identify anomalies that may have arisen from the concordance processes.

Problem postcodes

Concordances for two locations yielded results that were problematic: one around Tamworth in NSW and another around Weipa in Far North Queensland. In NSW, six SLAs were amalgamated in 2005 into Tamworth Regional (A) – Part A and Tamworth Regional (A) – Part B. This effectively changed the estimation method for SLAs affected by "pooling" a number of postcodes into the new (larger) entity and then redistributing (an averaged) proportion back to the smaller (2001) SLA. As a result, the individual characteristics of the smaller areas were lost as this data was merged with that of adjoining areas. In Far North Queensland, the number of SLAs was increased creating the reverse situation – that is, the "pooling" of earlier years was reduced. Both situations create an inconsistency with earlier estimates.

A test was performed to assess these differences by using the 2004 concordances to compile estimates of 2004–05 data and comparing this with that obtained using 2005 concordances. This test suggested that the effect of pooling was significant – especially where pooling occurred across urban and non-urban postcodes. Some consideration was given to using the 2004 concordances to amend the estimates for each postcode but, due to the interlocking nature of concordances, it was not feasible to apply the 2004 concordance to a group of postcodes in isolation without affecting surrounding estimates. On the other hand, relying entirely on the 2005 concordance estimates creates significant distortions for a number of SLAs particularly in the context of time series estimates.

Consequently, for a number of SLAs the estimates using 2005 concordances have been replaced with estimates that use the 2004 concordances. Note that these estimates still reflect 2004–05 postcode data, but the methodology has been altered to maintain consistency with earlier years. The small net changes in the totals caused by these changes are reflected in the number of "unknowns" in each state.

The need to make these adjustments will not continue beyond this year's estimates. Next year, the BITRE proposes to change the geography to ASGC 2006 and has engaged the ABS to construct single step concordances for all years covered in the database. These will create new estimates from the original postcode data, although estimates will not be made for 2001 SLAs that are not reflected in ASGC 2006.

The following table shows the difference between using the 2004 concordance compared with using the 2005 concordance for SLAs affected by boundary changes. The first three columns are the numbers in the final database.

Effects of methodological differences in estimates for selected postcodes – significantly affected SLAs

		using 2004 concordances			*using 2005 concordances*		
2001 SLA Code	*2001 SLA Name*	*Non-taxables 2004–05*	*Number of Taxable Individuals 2004–05*	*Aggregate Real Taxable Income 2004–05 (2006–07$)*	*Non-taxables 2004–05*	*Number of Taxable Individuals 2004–05*	*Aggregate Real Taxable Income 2004–05 (2006–07$)*
Northern NSW							
16000	Nundle (A)	149	569	24 273 633	165	476	18 463 790
16301	Parry (A) – Pt A	695	2 314	90 230 194	671	2 507	105 859 965
16304	Parry (A) – Pt B	928	2 680	110 124 651	895	2 549	101 892 436
17300	Tamworth (C)	4 287	16 307	696 144 827	4 312	16 117	680 597 062
10400	Barraba (A)	299	721	24 535 404	287	801	30 419 514
15100	Manilla (A)	419	1 015	36 109 823	410	1 186	46 034 533
Far North Queensland							
32501	Cook (S) (excl.Weipa)	1 252	2 781	125 210 159	1 019	2 258	102 148 689
32504	Cook (S) – Weipa only	163	731	41 753 597	191	858	49 023 854

References:

Australian Bureau of Statistics 2001, Australian Standard Geographical Classification 2001 (cat. no. 1216.0), ABS, Canberra, viewed November 2007 at http://www.abs.gov.au/AUSSTATS/abs@.nsf/66f306f503e529a5ca25697e0017661f/a3658d8f0ad7a9b6ca256ad4007f1c42!OpenDocument.

Australian Bureau of Statistics 2007, Consumer Price Index (cat. no. 6401.0), ABS, Canberra, viewed November 2007 at http://www.abs.gov.au/AUSSTATS/abs@.nsf/allprimarymainfeatures/0DECB9F37737E66FCA25737D00222523?opendocument.

Australian Taxation Office 2007, Taxation Statistics 2004–05: A summary of income tax returns for the 2004–05 income year and other reported tax information for the 2005–06 financial year, viewed November 2007 at www.ato.gov.au/corporate/content.asp?doc=/content/81183.htm&mnu=38022&mfp=001.

Bureau of Transport and Regional Economics 2005, Focus on Regions 3: Taxable Income, Information Paper 54, Bureau of Transport and Regional Economics, Canberra.

Reserve Bank of Australia 2007, Prices and Output (G Tables), Consumer Price Index (G2), viewed November 2007 at http://www.rba.gov.au/Statistics/Bulletin/index.html.