Australian National University

Crawford School of Public Policy

TTPI

Tax and Transfer Policy Institute

The ATO Longitudinal Information Files (ALife): Individuals - A new dataset for public policy research

TTPI - Working Paper 13/2021
July 2021

Thomas Abhayaratna
Crawford School of Public Policy
The Australian National University

Andrew Carter
Crawford School of Public Policy
The Australian National University

Shane Johnson
Crawford School of Public Policy
The Australian National University

Abstract

The Australian Taxation Office Longitudinal Information Files: Individuals (*ALife*: *Individuals*), is one of the most comprehensive tax administrative datasets in the world. The *ALife*: *Individuals* dataset, which currently covers the period 1990-91 to 2017-18, is based on a 10 per cent longitudinal sample of administrative unit-record personal income tax data. This new, high quality, longitudinal, de-identified, research-ready dataset is available to approved researchers through secure environments that safeguard taxpayers' information. The availability of *ALife: Individuals* opens exciting new possibilities for public policy research and evaluation that will improve understanding of taxpayer behavior and support policy development and its administration.

Keywords: personal taxation, longitudinal data

** The authors worked in the Australian Taxation Office to develop the Australian Taxation Office Longitudinal Information Files: Individuals (ALife: Individuals). They wish to acknowledge the contribution of other members of the team who assisted in developing and curating the dataset, including current and past colleagues of the Tax Policy Research and Analysis and Taxation Statistics Teams. Shane Johnson would like to acknowledge the support of the Sir Roland Wilson Foundation. We would like to thank Robert Breunig from the Tax and Transfer Policy Institute and Bruce Bastian for helpful comments. All findings, opinions and conclusions are those of the authors and do not represent the views of the Australian Government or any of its agencies. The proposal for this research was approved by the Australian National University's Human Research Ethics Committee, protocol number 2015/477.*

Tax and Transfer Policy Institute

Crawford School of Public Policy

College of Asia and the Pacific

+61 2 6125 9318

tax.policy@anu.edu.au

The Australian National University

Canberra ACT 0200 Australia

www.anu.edu.au

The Tax and Transfer Policy Institute (TTPI) is an independent policy institute that was established in 2013 with seed funding from the federal government. It is supported by the Crawford School of Public Policy of the Australian National University.

TTPI contributes to public policy by improving understanding, building the evidence base, and promoting the study, discussion and debate of the economic and social impacts of the tax and transfer system.

The Crawford School of Public Policy is the Australian National University's public policy school, serving and influencing Australia, Asia and the Pacific through advanced policy research, graduate and executive education, and policy impact.

Data survey

The ATO Longitudinal Information Files (*ALife*): Individuals – A New Dataset for Public Policy Research

Thomas Abhayaratna†, Andrew Carter†, Shane Johnson†

† Crawford School of Public Policy, Australian National University

1. Introduction

Recognising the limitations of the existing available administrative taxation data and growing demand to ensure the efficiency and effectiveness of government programs and service delivery, the Australian Taxation Office (ATO) embarked on the development of the ATO Longitudinal Information Files (*ALife*) to improve access to individual-level administrative data.

In 2018 the ATO released the first instalment of *ALife, ALife: Individuals* (from here referred to as *ALife*), a 10 percent sample of annual longitudinally linked individual level personal income tax records.[1] New waves of *ALife* are released annually, early each calendar year, with a two-year lag to allow for the lodgment cycle in tax returns. The current release, *ALife* 2018, released in early 2021, provides unit-record personal income tax data over the period 1990-91 through to 2017-18 and individually linked superannuation records from 1996-97 to 2017-18.[2]

Since the inception of the tax system, Australian taxation data has been made publicly available in various forms. Initially, aggregate taxation statistics were incorporated into the appendices of the Taxation Commissioner's Annual Report to Parliament, and from 1959-60 annual reports have been supplemented by the publication of Taxation Statistics.[3] The content in Taxation Statistics has also grown significantly over time. One significant inclusion has been a small sample of confidentialised, individual unit-record data that was first released with *Taxation Statistics 2003-04*.[4]

The availability of a small confidentialised individual unit-record dataset was a significant step forward in providing data access, and spurred growth in policy-relevant research. However, the dataset has several limitations.[5] Only a small number of variables are provided, limiting possible areas of research. In

---

1 *ALife: Individuals* is currently being used as the base to develop additional *ALife* modules, including: *ALife: Family Links* file to enable household and intergenerational analysis, and *ALife: Linked Employee Employer Data*, a linked personal income tax and employer payroll dataset.
2 Polidano et al. (2020) provide further information about the linked superannuation records.
3 The first five annual reports were made by the Commissioner of Land Tax, covering the period 1911–1916. Subsequent reports were made by the Commissioner of Taxation. Up until 1954-55 annual reports were not produced for each year.
4 For income years 2003-04 to 2010-11, files containing a one percent sample of records were made available. From 2011–12 files containing a two percent sample of records have been available.
5 Recent use of the sample file includes an examination of the progressivity of the Australian personal tax system following the introduction of a New Tax System (Tran & Zakariyya, 2020).

addition, the datasets are heavily 'perturbed', whereby measurement error is added to income variables which limits the use of standard techniques to answer straightforward questions, such as those that rely on discontinuity design, and bunching analysis (Gong & Gao, 2018). In addition, the data are cross sectional and therefore techniques such as fixed effects, which could be useful given the limited demographic information available in administrative tax records, cannot be utilised. Furthermore, fundamentally dynamic research questions such as income mobility cannot be addressed with cross-sectional data.

In recent years, we have seen greater demand for data. In part, this increased demand reflects increased emphasis on a desire for evidence-informed policy. This increased demand for data is supported by greater computing power, falling storage costs and improved analytical techniques (Productivity Commission, 2017). At the same time, we have also seen increased expectations around the availability and use of the vast amounts of data collected by government (Productivity Commission, 2017; Henry et al., 2010).

ALife provides a new dataset of individual unit-record personal income tax return data. This new, longitudinal, de-identified dataset is available to approved users through monitored, secure environments that safeguard the confidentiality and privacy of the taxpayers' information.

The remainder of this paper is structured as follows. Section 2 provides an overview of the Australian personal income tax system. Section 3 briefly outlines why *ALife* was created. Section 4 provides information on the construction of the dataset, supporting materials for researchers, and measures to protect taxpayer privacy. Section 5 provides information on access arrangements. Section 6 describes planned extensions and possible enhancements that could improve the potential of the dataset for research and policy analysis.

2. Australia's personal income tax system

Australia first introduced a federal income tax in 1915 to finance involvement in the First World War. The federal income tax was modelled on the income tax systems that applied in the Australian states at the time and the comprehensive income tax system that applied in the United States (see Reinhart and Steel (2006) for a brief history of the Australian taxation system and Tilley (2020) for a detailed history of the early tax policy developments following Australia's federation). While Australia's personal income tax system has gone through many changes since its inception in 1915, the basic principle of comprehensive taxation has remained.

Australia's current personal income tax system shares many similarities to the personal income tax systems adopted across Western economies. The Australian personal tax system is based on the individual, although family circumstances are recognised for a number of elements; including the Medicare Levy (and surcharge) the dependent tax offset, and the Family Tax Benefit (a means tested

payment for families with children that could be claimed as a tax offset until 2008).[6]

Australian residents are required to report their worldwide assessable income, which includes wages and salaries; pensions and allowances; interest; dividends; capital gains (on realisation); net rental income and income from businesses, partnerships and trusts. Deductions are allowed for the costs incurred in earning assessable income, as well as other expenses including charitable donations, costs of managing tax affairs and union membership.

A progressive rate structure has also been a core feature of the Australian personal tax system since its inception. While the top marginal tax rate has decreased over the past 50 years, it has remained relatively stable since 1990-91 when it was reduced to 47 percent and since the 2017-18 income year has been 45 percent. From 2017-18, $18,200 could effectively be earned before tax applied as a result of the generous tax-free threshold, which has increased from $5,250 in 1990-91.[7]

The personal income tax system has also been extensively used to support other policy objectives. Several tax offsets, which act to reduce a taxpayer's tax liability, have been available. Most notably these offsets include the Low Income Tax Offset, which reduces tax paid by low income individuals, and the Seniors and Pensioners Tax Offset, which aims to reduce the tax liability for individuals of Age Pension age or receiving a pension through the social security system.

The personal tax system is also used to encourage individuals and families to take up private health insurance. The Medicare Levy Surcharge applies to individuals without private health insurance and for whom their relevant income exceeds the applicable threshold (based on the family status of the taxpayer).

In addition, the personal tax system is also used to support Australia's higher education income contingent loan schemes. Under the schemes, Commonwealth provided loans for higher education fees are repaid through the tax system based on the individual's income.

The personal tax system has also been used to support fiscal stimulus, including around 8.4 million tax bonus payments, with a total value of around $7.3 billion in April and May 2008 as part of the government's response to the global financial crisis. Temporary levies have also been imposed and administered through the tax system, including the 2011 Flood Levy which helped to fund the reconstruction efforts in Queensland following the 2010-11 floods and the Temporary Budget Repair Levy, which applied from 2014 to 2017.

---

[6] Since the 2009 income year, the Family Tax Benefit has been administered through the social security system.
[7] The Low Income Tax Offset extends the tax-free threshold for eligible low-income earners. In 2018, the effective tax-free threshold (including the Low Income Tax Offset) was around $20,542.

2. Why *ALife* was created

The availability of administrative data has expanded in many countries over recent decades. The potential benefits of these data for research are significant, allowing for deeper analysis and understanding of tax and social security systems and supporting improved design. However, Australia has lagged comparable countries, particularly for access and use of administrative, longitudinal data.

Most existing longitudinal data initiatives in Australia are used by researchers to examine health outcomes.[8] One notable exception is the Household, Income and Labour Dynamics in Australia (HILDA) Survey which, as its name suggests, is survey based and provides a range of income and social indicators. While surveys have their advantages, particularly for researching rich behavioral issues with detailed demographic and attitudinal variables, they can be expensive to conduct, rely on personal observations, have relatively small sample sizes and can be affected by sample attrition issues. High-quality administrative data can therefore complement survey data, particularly for public policy evaluation (Card et al., 2010; Connelly et al., 2016). Research published in leading economic journals using administrative data have also increased significantly with more administrative data being made available to researchers (Chetty, 2012). The potential for linking longitudinal survey and administrative data will provide more possibilities for public policy research and evaluation.

The Nordic countries have long-standing experience in curating and making administrative data available for research (United Nations, 2007). Denmark, Finland, Norway and Sweden have all used government register data for research since the late 1960s and early 1970s and Denmark, in 1981, was the first country to produce a national census using only data from administrative sources (United Nations, 2007). Importantly, the development and use of these administrative data registers have supported the improvement of social security and taxation systems (Statistics Finland, 2004).

Canada's statistics agency has released the Longitudinal Administrative Databank since 1994, created primarily from income tax and welfare benefits data. The U.S., U.K. and New Zealand have greatly expanded researcher access to administrative data in recent years (Almunia M. , Harju, Kotakorpi, Tukiainen, & Verho, 2019).

With this global trend, governments are competing for the attention of the world's best minds. Academics will naturally gravitate towards the richest sources of information for their research. Countries with limited access to data are therefore at risk of losing out on the (free) contribution that independent research can make to the advancement of national public policy and administration. Furthermore, policymakers in Australia, a country with a distinct policy setting, society and legal system, cannot

[8] See for example the Population Health Research Network (PHRN), an Australian wide national data linkage network which supports researchers to access linked population and health data (https://www.phrn.org.au).

necessarily rely on evidence drawn from other countries to inform decision-making. Increasing access to administrative data provides a robust foundation for policymaking and promotes domestic research capability and output.

The Australia's Future Tax System Review highlighted the importance of expanded access to confidentialised data to better understand the impacts of the tax and transfer system and support policy development, recommending confidentialised tax unit records be made freely available for analysis and research (Henry et al., 2010).[9] In June 2013 the ATO began a project to examine ways to make longitudinal data available to researchers to deepen the evidence base to inform public policy development, enhance understanding of behaviour and improve administration. From the outset, the project was underpinned by four key principles:

- make a high-quality dataset available for research;
- ensure the data is representative of the tax paying population;
- make the data as easy to use as possible; and
- ensure that taxpayer information remains private and confidential.

3. Construction

3.1. Sampling and retention

ALife follows individuals over time, tracked using their unique client identification number. The initial *ALife* file, *ALife* 2016, was based on a random 10 percent sample of the total population of individuals on the ATO's 2016 client register. The client register, which is actively maintained since at least 1980, includes all individuals who are, or have been, registered by the ATO. The most common reason an individual appears on the client register is when they first apply for a tax file number (TFN). The client register includes all active clients (those still expected to engage with the ATO including temporary visa holders) and inactive clients (those that are no longer expected to engage with the ATO such as deceased individuals or temporary visa holders that have left Australia). The client register also accounts for individuals who have been reissued or changed their TFN.[10] Individuals included on the client register may not have ever lodged a personal income tax return.

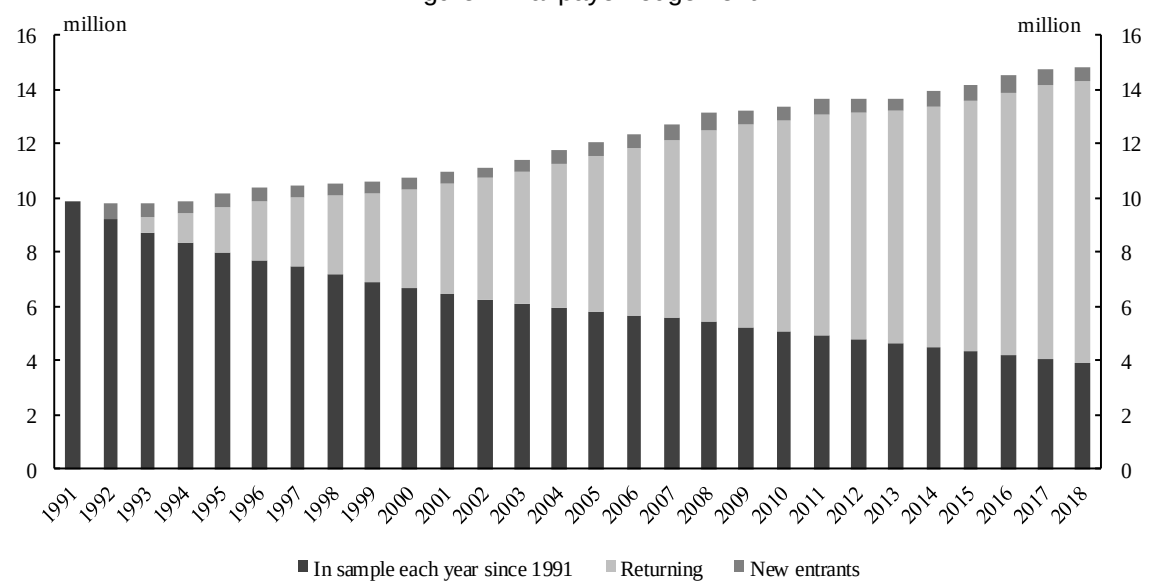
To draw the random sample from the client register, everyone on the register in 2016 was allocated a permanent number between 0 and 1, and all individuals with a number less than 0.1 were included in the sample. Under this approach, each person on the register had a 10 percent probability of being drawn into the sample. Those selected remain in the file until they reach the maximum cut-off age of currently

[9] See recommendation 133, Henry et al. (2010 p.722).
[10] A small number of tax filers identified by the ATO as 'protected' are excluded from *ALife*.

set at 93 from the year 1999-2000.[11] Each year a 10 percent sample of all new individuals who are added to the client register in the income year is randomly selected and added to the main data file, thus ensuring the sample remains a 10 percent representative sample of the register. As tax lodgment and amendment activity occur on an ongoing basis, a fixed date, the first Monday in November each year, is used as the cut-off to sample individuals and extract data. This approach is consistent with the date used in the Taxation Statistics publications (Australian Taxation Office, 2021).

Administrative longitudinal datasets are less susceptible to sample attrition compared to survey-based longitudinal datasets. *ALife* has a high level of retention. On average, around 96.5 percent of tax filers who lodge in a given year also lodge in the subsequent income year. After 10 years, around 65 percent of respondents who were in the original sample remain and lodged in each year, this falls to around 50 percent after 20 years and around 40 percent after 30 years in the current 2018 wave (Figure 1). In addition, many tax filers who drop out of the data return to the sample, lodging a tax return in a future income year. There are several reasons for non-lodgment in a given year, most commonly individuals are not required to lodge a return when their taxable income is below the tax-free threshold and no tax has been paid on their taxable income.[12]

Figure 1: Taxpayer lodgement

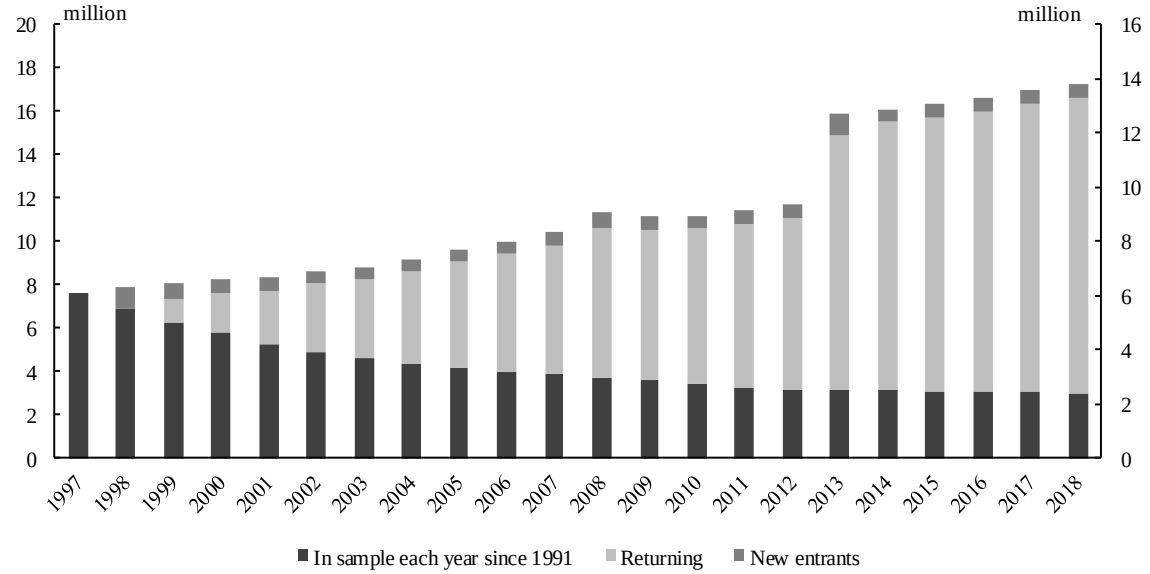


Source: Author's calculations.

[11] The maximum cut-off age is increased by one year every five years.
[12] Section 7 below, discusses future possible enhancements to *ALife:* individuals, including providing data on non-lodgers based on third party reporting.

Administrative data are, however, not without problems. Policy and administrative changes can have significant implications on who is captured in administrative data and on what data are collected. An example of this can be seen in the linked *ALife* superannuation data, where prior to the 2012-13 income year superannuation funds were only required to lodge a Member Contribution Statements (MCS) if a contribution was made into the member's account for a given income year. As such, many 'inactive' accounts are not included in *ALife* prior to 2013 (Polidano, et al., 2020).[13] Since 1 July 2012, reporting obligations changed meaning superannuation funds were required to lodge an MCS for all their members, including those for whom no contributions were received. This change resulted in a structural break in *ALife* superannuation data in 2013, when around 3.4 million 'inactive' accounts where included in the sampling population (Figure 2).

Figure 2: Superannuation lodgement



Source: Author's calculations.

3.2. Supporting usability

Significant effort has also gone into making *ALife* as research-ready and user friendly as possible. Ensuring the data are representative and high quality is a central focus in the construction and curation of the dataset. This included testing the data by micro-simulating elements of the tax return using standalone modules and the comprehensive AusTaxSim model (Johnson, 2021).

To support usability, a clear naming convention was developed for all variables within *ALife*. The naming convention is based on location of a given variable in the tax return (for example, i_ income variables, is_ income variables in the supplementary return, d_ deductions and so forth) and how the

[13] Some superannuation fund managers voluntarily lodged MCS for inactive accounts.

variable is used (see Appendix A).[14] This approach allows users to utilise past tax returns as a conceptual guide to the dataset.[15] The naming convention also includes abbreviations and mnemonics to shorten the length of variables for statistical software with maximum character length limits.

ALife is supported by an online interactive manual, which includes a data dictionary with summary information for each variable included in the dataset and annotated tax returns (which includes the variable name that corresponds to a particular field on the tax return form) as well as tax return instructions.[16] A forum to allow users to discuss, ask and answer questions relating to the dataset is currently under development.

3.3. Protecting privacy

ALife has been designed to provide data with as much detail as possible, while safeguarding privacy and security.[17] The project uses the 'Five Safes' framework that was originally developed by the UK Office for National Statistics to manage disclosure risk (Ritchie, 2008).[18] The personal information of individual tax filers is protected through several measures. These include the removal of direct and indirect identifiers from the dataset such as, names and tax file numbers. An additional safeguard is provided by the 10 percent sample, which further reduces the probability of re-identification.[19]

ALife strikes a careful balance between treating variables to protect the privacy of individual tax filers and ensuring access to a robust, useable dataset.[20] To reduce the risk of spontaneous re-identification, a small number of other variables have been treated. Eligible termination benefit payments in each year are top-coded, as these payments are typically made to departing chief executive officers of major listed Australian companies and often publicly reported. In these cases, the individual's tax liability, and related variables, have been recalculated. Election expense deductions are aggregated with 'other expenses' to reduce the risk of spontaneous recognition of individuals running for public office or making donations that are listed on public registers. An individuals' date of birth is replaced with their age as of June 30 (presented as an integer) of the tax year – the age typically required for age dependent means tests.[21]

---

[14] Some variables have moved within the tax return. Where this has been the case the variables location in the 2016 base year has been used as the default for the naming convention.
[15] Tax returns marked-up with variables names are available on the *ALife* website, at: https://ALife-research.app/research/forms/itr
[16] The online interactive manual can be accessed at: https://ALife-research.app.
[17] ALife complies with the Privacy Act 1988 (Cth) and the Taxation Administration Act 1953 (Cth).
[18] The Five Safes Framework is also used by other Australian Government data custodians, including the Australian Bureau of Statistics (2017) and the Australian Institute of Health and Welfare (2020).
[19] Privacy and security are further protected by granting access only to approved researchers that have committed to safeguard the data, accessing and storing the data in a high-security data laboratory, and vetting all requests for data output, this is discussed at Section 6 below.
[20] Independent assessments on the ALife project were also conducted – including a privacy impact assessment conducted by Salinger Privacy (https://alife-research.app/info/overview#Protecting%20privacy) and a quantitative risk analysis by the Australian Institute of Health and Welfare – to strengthen the privacy protections.
[21] Some public pension eligibility tests require more granular detail on age, to support this a binary indicator variable is included to indicate eligibility.

Further, as there are only a small number of tax filers over the age of 90, individuals above a maximum age are excluded from the dataset – this age is set at 93 from the year 1999-2000, and increases by one year for every five years. Information about the date a taxpayer partnered or separated during a year is converted into a flag indicator. Address information has been suppressed and taxpayer location information is provided only up to the corresponding Statistical Area Level 4.[22]

4. Information available within *ALife*: Individuals

The current release of *ALife*, *ALife* 2018, provides longitudinal, unit-record personal income tax and superannuation information for tax filers over the period 1990-91 to 2017-18. The *ALife* sample follows the tax returns of almost 1.5 million Australians in 2017-18.

Around 300 variables from the personal income tax return are included in the *ALife* dataset, in addition to over 100 superannuation variables from Super Member Contribution Statements (MCS) and Self-Managed Superannuation Fund (SMSF) annual returns. *ALife* includes several derived variables that intend to increase research potential including, for example, an indicator on whether the tax filer has a spouse (based on information from the tax return). Variables from the personal income tax return cover all information required to assess an individual's income tax liability and include income, allowable deductions, tax offsets, higher education income contingent loans, Medicare Levy and Surcharge and any credits or refundable tax offsets (Appendix B provides a summary of the various elements included in the dataset, and their relationship to the personal tax return).

ALife provides the most comprehensive view of the components of individual income available in any Australian dataset. Access to wage, investment, rental and capital gain incomes, and tax paid allows researchers to study the distributional qualities of net income across each component. The richness of the data may also allow it to be used as a benchmark for other data sources.

Tax filers' demographic information in *ALife* is somewhat limited as detailed information is not collected through the tax system. Demographic information included in the dataset includes age at 30 June of the income year, gender, residential geographic location mapped to the Statistical Area Level 4, occupation and whether the tax filer is a non-resident for tax purposes.[23]

A selection of summary information from a subset of the *ALife* 2018 variables is provided in Table 1 below. This table compares key statistics for the sample file against the population of administrative tax data (a summary across all years provided at Appendix C). Notably the *ALife* 10 percent sample is representative of the tax-filing population. In the 2018 wave, the full *ALife* tax filer population included around 14.8 million tax filers, with around 1.48 million tax filers in the 10 percent sample.

---

[22] At the Statistical Area Level 4 level, when comparing any two years, there are tens to hundreds of individuals moving between any two regions and as such the risk of re-identification is low.

[23] To reduce possible reidentification risks, a tax filer's identified gender is fixed over the panel.

Table 1: ALife Summary Statistics of Selected Variables
Count, Mean and Standard Deviation, 2017-18

	Total		Male		Females	
Variable	ALife 10 per cent	ALife population	ALife 10 per cent	ALife population	ALife 10 per cent	ALife population
Observations	1,483,492	14,840,381	762,035	7,629,874	721,457	7,208,665
Age	43.1	43.1	43.2	43.2	43.1	43.1
	(16.23)	(16.22)	(16.2)	(16.19)	(16.25)	(16.25)
Salary and wages	47,781.3	47,781.0	56,418.8	56,418.9	38,668.4	38,649.5
	(58,117.98)	(61,477.29)	(68,980.74)	(73,962.13)	(41,926.87)	(42,752.65)
Total income	63,395.3	63,316.4	74,613.8	74,496.9	51,559.5	51,496.3
	(195,828.1)	(214,100.6)	(243,429.4)	(236,570.1)	(126,485.2)	(186,702.9)
Total deductions	2,510.7	2,545.0	3,067.4	3,122.5	1,923.2	1,934.3
	(19,759.78)	(81,047.11)	(15,512.83)	(80,697.14)	(23,411.44)	(81,421.7)
Taxable income	60,772.3	60,656.1	71,411.2	71,233.2	49,548.1	49,474.1
	(190,876.8)	(190,253.1)	(239,123.7)	(213,793)	(119,512.1)	(160,911.2)

Source: Author's calculations.
Note: Standard deviations in parenthesis.

5. Access arrangements

Access to *ALife* is limited to researchers from approved organisations. A head agreement with the ATO must be in place before any of an organisation's researchers can apply to access the file. The *ALife* governance arrangements are designed to ensure that organisations put in place adequate controls and oversight for their researchers, as both the organisation and researcher can face consequences for breaches of a research protocol.

Currently, there are 21 approved Australian universities with a head agreement with the ATO, in addition to a smaller number of approved government agencies. For a specific project to be considered by the ATO, researchers are required to provide a description of their proposed research and methodology, a public interest statement and are required to disclose any real or perceived conflicts of interest. In addition, researchers are also required to provide evidence of ethics approval from a registered Australian Human Research Ethics Committee. The ATO considers the application based on the feasibility of research and any additional costs (staffing and monetary) that may be incurred. Researchers are also required to complete customised training and a short quiz to test the understanding of their responsibilities prior to accessing the dataset.

Approved researchers generally access *ALife* through a secure connection to a remote-access data research laboratory operated by an approved independent provider. Currently, access is available through the Secure Unified Research Environment (SURE), operated by the Sax Institute. Researchers' sessions within the SURE environment are available to the ATO for monitoring and compliance auditing. Researchers looking to take aggregate data or results out of the environment or bring other data into the environment are required to apply to the ATO for approval. The ATO reviews all aggregate data and empirical research output before it is released from the secure environment to ensure the data cannot be

used to re-identify individuals and is consistent with the approved research project.[24] The ATO intends to list *ALife* research output on the A*Life* website.

The ATO's approach is similar to the Nordic countries who have a long history in providing access to unit-record level data. For example, Statistics Sweden provides access to approved projects through the Microdata Online Access (MONA) delivery system with projects undergoing a confidentiality assessment by Statistics Sweden, and ethical approval (for access to sensitive personal data) (Statistics Sweden, 2021). Researchers are charged for the cost of providing the data, typically between SEK 30,000 and SEK 60,000 (between AUD 4,700 and AUD 9,400) (Statistics Sweden, 2021). Similarly, Statistics Denmark only provide access to unit record data for approved projects, undertaken by a researcher affiliated by a research environment pre-approved by Statistics Denmark (Statistics Denmark, 2021). Statistics Denmark also charge for the cost of extraction (in July 2018 this was DKK 1,150 per hour, or around AUD 250) plus storage costs for large projects (Statistics Denmark, 2018).

6. Dataset extensions and possible enhancements

ALife provides a basis that could be used as a model to support greater access to other personal level administrative data. A household identifier is currently being developed that will enable intergenerational analysis, while a linked longitudinal employee and employer module is also being developed.

Adding data on non-lodgers, that is those individuals who don't lodge a personal income tax return (including, the young with little income, and senior Australians) is also being developed. This could include income related data from third parties, such as that used to pre-fill electronic income tax return forms, such as pay-as-you-go (PAYG) withholding income tax summaries, investment and interest income data and government payments data.

Work is also underway to develop a small 'synthetic' version of the file, which would allow users to scope out possible research ideas and draft code outside the secure environment. If some statistical integrity of the data can be preserved in the synthetic version, it could provide a means for some types of analysis to be performed outside the environment without incurring a financial cost to access the secure environment or requiring ethics approval for the analysis.[25]

The *ALife* project also presents opportunities to link ATO administrative data to other datasets in the future. For example, social security data and *ALife* data could potentially be combined to provide improved insight on interactions between the tax and transfer system, health information could be added to better understand determinants of health outcomes, and education data. It is also possible that the data

[24] Breaches of the terms of use by individuals or institutions could result in termination of access. Some breaches, such as attempting to re-identify individuals from the dataset or unauthorised disclosures of information, carry criminal penalties.

[25] Synthetic data could also be used to address confidentiality concerns were datasets are linked. For example, the United States Census Bureau's Survey of Income and Program Participation Synthetic Beta (SSB) integrates individual-level micro-data from a household survey with administrative tax and benefit data (Benedetto, Stinson, & Abowd, 1993).

underpinning the *ALife* project could be linked to survey data such as the Household, Income and Labour Dynamics in Australia (HILDA) Survey or support the creation and validation of official statistics.

7. Conclusions

ALife provides researchers with access to one of the world's most comprehensive personal income tax data resources. Addressing a growing need for access to high quality administrative data, the *ALife* platform, and its first module *ALife*: *Individuals*, provides researchers with access to unit-record level personal income tax and superannuation data. This file presents significant opportunities for new public policy research, and over time, will support greater evidence-informed policy.

Importantly, the curation and release of *ALife* in line with Five Safes framework ensures detailed administrative data are available to researchers while preserving the privacy of taxpayers.

The *ALife* project will continue to be developed to provide researchers with access to additional useful material, including a householder identifier and data for non-lodgers, while opportunities to link with other administrative datasets will be explored. These initiatives will expand the use of the *ALife* platform to support greater policy analysis and research and in turn support better evidence informed policy making.

Correspondence

Correspondence on the ATO Longitudinal Information Files should be directed to: ALife@ato.gov.au.
Further information on the ATO Longitudinal Information Files can be found at: ALife-research.app/
Corresponding author: Shane.Johnson@anu.edu.au.

References

Almunia, M., Harju, J., Kotakorpi, K., Tukiainen, J., & Verho, J. (2019). Expanding access to administrative data: the case of tax authorities in Finland and the UK. *International Tax and Public Finance, **26***(3), 661-676.

Australian Bureau of Statistics (ABS). (2017). Managing the Risk of Disclosure: The Five Safes Framework, In ABS, *ABS Cat. No. 1160 ABS Confidentiality Series.* Retrieved January 12, 2021, from https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1160.0Main%20Features4Aug%202017

Australian Institute of Health and Welfare. (2020). *Data Governance framework.* Retrieved January 21, 2021, from https://www.aihw.gov.au/about-our-data/data-governance

Australian Taxation Office. (2021). *Taxation Statistics 2018-19*. Retrieved June 17, 2021, from Taxation Statistics 2018–19: https://www.ato.gov.au/About-ATO/Research-and-statistics/In-detail/Taxation-statistics/Taxation-statistics-2018-19/

Benedetto, G., Stinson, M., & Abowd, J. (1993, April). The Creation and Use of the SIPP Synthetic Beta. Retrieved January 27, 2021, from https://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf

Card, D., Chetty, R., Feldstein, M., & Saez, E. (2010). Expanding Access to Administrative Data for Research in the United States, *American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas.*

Chetty, R. (2012). *Time Trends in the Use of Administrative Data for Empirical Research.* Retrieved January 21, 2021, from Presented at the NBER Summer Institute: http://www.rajchetty.com/chettyfiles/admin_data_trends.pdf

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research, 59*, 1–12.

Gong, X., & Gao, J. (2018). Nonparametric Kernel Estimation of the Impact of Tax Policy on the Demand for Private Health Insurance in Australia. *Australian & New Zealand Journal of Statistics, **60***(3), 374-393.

Henry, K., Harmer, J., Piggott, J., Ridout, H., & Smith, G. (2010). *Australia's Future Tax System, Report to the Treasurer, Volume 2, Detailed Analysis.* Canberra: Australian Government, The Treasury.

Johnson, S. (2021). The AusTaxSim Model – A Static Tax Microsimulation model for use with ALife: Individuals. TTPI Working Paper.

Polidano, C., Carter, A., Chan, M., Chigavazira, A., To, H., Holland, J., Nguyen, S., Vu, H., Wilkins, R. (2020). The ATO Longitudinal Information Files (ALife): A New Resource for Retirement Policy Research. *Australian Economic Review, **53***(3), 429-449.

Productivity Commission. (2017). Data Availability and Use. *Report No. 82*. Canberra: Commonwealth of Australia.

Reinhardt, S., & Steel, L. (2006). A Brief History of Australia's Tax System. *Economic Round-up,* Australian Government, The Treasury.

Ritchie, F. (2008). Secure access to confidential microdata: four years of the Virtual Microdata Laboratory. *Economic and Labour Market Review, **2***(5), 29-34.

Statistics Denmark. (2018). *Step by step practices for national register based research projects at Statistics Denmark.* Retrieved January 29, 2021, from https://www.dst.dk/-/media/Kontorer/13-Forskning-og-Metode/Step-by-step-

procedures-for-researchers-access-to-Microdata_082018.pdf?la=en

Statistics Denmark. (2021). *Data for research*. Retrieved January 29, 2021, from https://www.dst.dk/en/TilSalg/Forskningsservice

Statistics Finland. (2004). *Use of Registers and Administrative Data Sources for Statistical Purposes - Best Practices of Statistics Finland*. Helsinki: Valopaino. Retrieved January 11, 2021, from http://unstats.un.org/unsd/EconStatKB/Attachment172.aspx?AttachmentType=1

Statistics Sweden. (2021). *How to order microdata from Statistics Sweden*. Retrieved January 29, 2021, from Statistics Sweden: https://www.scb.se/en/services/guidance-for-researchers-and-universities/

Tilley, P. (2020). Early federation reviews and 1942 income tax unification. *Tax and Transfer Policy Institute Working paper 11/2020*.

Tran, C., & Zakariyya, N. (2020). Tax Progressivity in Australia: Facts, Measurements and Estimates. *Economic Record*, **97**(316), 45-77.

United Nations. (2007). Register-based statistics in the Nordic countries – Review of best practices with focus on population and social statistics. New York: United Nations.

Appendix A. *ALife* variable naming conventions

Mnemonic	Description
a_	adjustments
b_	business
c_	client information
cd_	code
cr_	credit
d_	deduction variable
dc_	deductions calculated
div293_	super variable related to Division 293
dn_	deductions information
ds_	deductions supplement
dsc_	deductions supplement calculated
dsn_	deductions supplement information
f_	flag variable
h_	HECS / higher education
hc_	HECS calculated
i_	income variable
ic_	income calculated
in_	income information
is_	income supplement
isc_	income supplement calculated
isn_	income supplement information
it_	income tests
itc_	income test calculated
l_	losses
ln	losses information
ml_	Medicare levy
mls_	Medicare levy surcharge
o_	offset variable
oc_	offset calculated
on_	offset information
onra_	offest non-refundable available
onru_	offest non-refundable used
or_	offset refundable
osc_	offset supplement calculated
pt_	parternships and trusts
sb_	super balance
sc_	super contribution
scc_	super co-contribution
sd_	super drawdown
sn_	other super information
sp_	spouse information
st_	super transfer
tc_	tax calculated
tn_	tax information variable
tw_	tax withheld

Abbreviation	Description
accum	accumulation
annuities	Australian annuities and super income streams
apra	Australian Prudential Regulation Authority
aust	Australia
bal	balance
bus	business
calcd	calculation
cfc	controlled foreign company
cg	captial gains
cont	contribution
cr / crs	credit
db	defined benefits
ded	deducted
defer	deferred
depend	dependent
disc	discount
dist	distribution / distriubted
div	dividend
edu	education
elgbl	eligible
empl	employer / employment
ess	employee share schemes
eto	entreprenures tax offset
etp	employment termination payments
exp	expense
fam	family
fmd	farm management deposits
fmis	forestry managed investment scheme
frank	franked
fsi	foreign source income
gov	government
inc	income
inelgbl	ineligible
insure	insurance
invest	investment
lito	Low Income Tax Offset
loss	losses
lsum	lump sum
mawto	Mature Age Worker Tax Offset
mis	managed investment scheme
npp	non primary production
nras	national rental affordability scheme
ntnl	notional
nz	New Zealand
phi	private health insurance
pp	primary production
pps	prescribed payments system
prev	previously
prof	professional
pship	partnership
psi	personal services income
sato	Senior Australians Tax Offset
smsf	self managed super fund
sni	separate net income
super	superannuation
supp	supplement
txd	taxed
untxd	untaxed
upp	undeducted purchase price
vet	veteran
wre	work related expense

Appendix B. Elements of *ALife*: Individuals

Category	Variable
Demographic	• Age at 30 June* • Sex • Geographic location at SA 4* • Occupation • Whether the tax filer has a spouse* • Whether the tax filer is a resident for tax purposes*

Income (i_ is_)	• Salary and wages
	• Allowances, earnings, tips, directors' fees etc.
	• Lump sum payments
	• Eligible termination payments
	• Government allowances and payments (including)^:
	  ○ Jobseeker payment
	  ○ Newstart allowance
	  ○ youth allowance
	  ○ Austudy payment
	  ○ ABSTUDY living allowance
	  ○ parenting payment (partnered)
	  ○ partner allowance
	  ○ sickness allowance
	  ○ special benefit
	  ○ widow allowance
	  ○ farm household allowance
	• Government pensions (including)^:
	  ○ age pension
	  ○ bereavement allowance
	  ○ carer payment
	  ○ disability support pension
	  ○ parenting payment (single)
	  ○ age service pension
	  ○ income support supplement
	  ○ veteran payment
	  ○ invalidity service pension

	○ partner service pension • Superannuation pensions and annuities • Personal services income • Reportable fringe benefits • Investment income (including): ○ gross interest ○ dividends ○ net rent ○ managed investment trusts ○ realised capital gains • Business, partnership and trust income
Allowable deductions (d_ ds_)	• Work related expenses (including): ○ car expenses ○ travel expenses ○ uniform, clothing, laundry and dry cleaning ○ self-education expenses ○ tools, equipment and other assets (low value pool deduction) ○ other work-related deductions • Investment income deductions: ○ interest ○ dividends ○ other investment income • Gifts or donations • Cost of managing tax affairs • Personal superannuation contributions
Losses (l_)	• Tax losses of earlier years claimed in the current year
Taxable income (ic_)	• Income *less* allowable deductions *less* prior year losses claimed

Tax on taxable income (tc_)	• Gross tax on taxable income, based on the tax schedule
Tax offsets (on_)	• Spouse, child-housekeeper or housekeeper • Senior Australians and Pensioners • Australian Government allowances and payments • Private Health Insurance • Baby Bonus • Low income • Superannuation related offsets • Zone tax offset • Medical expenses
Net tax payable	• Gross tax *less* offsets
Private Health Insurance details	• Premium is eligible for the private health insurance rebate • The private health insurance rebate was received
Medicare levy and surcharge	• Number of dependent children and students • Days fully exempt from Medicare levy • Days half exempt from Medicare levy • Whether tax filer (and dependents) had private health insurance for the income year • Days exempt from Medicare levy surcharge
HECS-HELP and SFSS liability (hc_)	• Repayment income • HECS-HELP or SFSS liability • HECS-HELP or SFSS debt
Adjustments	• Tax filer was under 18 years old on 30 June of tax year • Month the tax filer was eligible for the part-year tax free threshold • Working holiday maker income
Tax credits and refundable offsets	• Imputation/franking credits

* Derived variable.
^ While these sources of income are included in *ALife*, they may not be separately identifiable.

Appendix C. Summary Statistics, selected variables

Year	Data set	Observations	Age	Salary and wages	Total income	Total deductions	Taxable income
2018	10 per cent	1,483,665	43	47,781	63,395	2,511	60,772
			(16)	(58,118)	(195,828)	(19,760)	(190,877)
	Population	14,840,381	43	47,786	63,323	2,545	60,662
			(16)	(61,479)	(214,111)	(81,052)	(190,262)
2017	10 per cent	1,473,750	43	45,816	60,855	2,514	58,206
			(16)	(57,096)	(144,883)	(13,510)	(141,673)
	Population	14,737,478	43	45,798	60,771	2,563	58,070
			(16)	(59,431)	(168,240)	(60,378)	(149,487)
2016	10 per cent	1,448,098	43	45,497	60,558	2,547	57,876
			(16)	(65,689)	(250,553)	(13,009)	(245,479)
	Population	14,485,416	43	45,452	60,345	2,557	57,648
			(16)	(62,157)	(173,974)	(17,911)	(169,322)
2015	10 per cent	1,419,009	43	44,724	59,529	2,644	56,734
			(16)	(59,502)	(127,379)	(59,622)	(134,162)
	Population	14,189,447	43	44,717	59,654	2,621	56,885
			(16)	(58,435)	(190,105)	(46,098)	(179,105)
2014	10 per cent	1,392,684	43	44,229	59,137	2,516	56,474
			(16)	(56,091)	(450,497)	(24,942)	(447,613)
	Population	13,928,794	43	44,234	58,846	2,500	56,193
			(16)	(58,956)	(198,989)	(22,025)	(194,692)
2013	10 per cent	1,365,847	43	43,676	57,370	2,423	54,802
			(16)	(54,277)	(147,090)	(19,564)	(132,577)
	Population	13,661,437	43	43,647	57,314	2,412	54,757
			(16)	(55,984)	(126,596)	(19,222)	(121,649)
2012	10 per cent	1,365,916	43	41,544	54,861	2,547	52,176
			(16)	(51,383)	(164,459)	(29,576)	(143,697)
	Population	13,665,367	43	41,512	54,728	2,518	52,063
			(16)	(52,202)	(119,994)	(20,012)	(112,589)
2011	10 per cent	1,359,954	42	39,011	51,869	2,505	49,201
			(16)	(49,998)	(116,041)	(84,814)	(139,587)
	Population	13,611,360	42	38,980	51,867	2,450	49,246
			(16)	(52,366)	(126,453)	(32,509)	(125,212)
2010	10 per cent	1,337,003	42	36,706	48,391	2,422	45,819
			(16)	(47,168)	(96,921)	(69,462)	(108,053)
	Population	13,379,083	42	36,673	48,436	2,359	45,931
			(16)	(48,621)	(144,590)	(27,714)	(143,129)
2009	10 per cent	1,319,283	42	35,580	47,454	3,008	44,316
			(16)	(58,301)	(387,507)	(496,355)	(161,222)
	Population	13,202,563	42	35,531	47,146	2,585	44,432
			(16)	(51,710)	(164,916)	(158,530)	(116,231)
2008	10 per cent	1,313,852	42	33,906	46,228	2,632	43,476
			(16)	(49,445)	(218,973)	(19,941)	(215,098)
	Population	13,146,048	42	33,913	46,221	2,644	43,456
			(16)	(104,088)	(217,515)	(84,230)	(197,317)
2007	10 per cent	1,266,481	42	31,875	45,025	2,849	42,063
			(16)	(46,362)	(117,574)	(19,340)	(113,219)
	Population	12,670,973	42	31,874	45,105	2,840	42,155
			(16)	(47,818)	(305,317)	(22,292)	(303,011)
2006	10 per cent	1,232,418	42	30,275	41,634	2,309	39,223
			(16)	(51,446)	(82,857)	(12,788)	(79,263)
	Population	12,331,140	42	30,234	41,670	2,312	39,258
			(16)	(42,231)	(85,967)	(14,969)	(81,363)
2005	10 per cent	1,203,105	42	28,838	39,400	1,995	37,308
			(16)	(36,191)	(73,519)	(14,411)	(69,090)
	Population	12,036,972	42	28,851	39,404	1,993	37,313
			(16)	(38,897)	(78,627)	(12,326)	(75,027)

Source: Author's calculations. Note: Standard deviations in parentheses.

Appendix C. Summary Statistics, selected variables (continued)

Year	Data set	Observations	Age	Salary and wages	Total income	Total deductions	Taxable income
2004	10 per cent	1,171,997	41	27,501	37,360	1,801	35,269
			(16)	(35,129)	(61,308)	(9,721)	(59,891)
	Population	11,730,164	41	27,492	37,365	1,807	35,266
			(16)	(34,818)	(69,352)	(12,740)	(66,591)
2003	10 per cent	1,138,673	41	26,415	35,596	1,660	33,680
			(16)	(33,529)	(66,041)	(8,431)	(65,167)
	Population	11,398,499	41	26,411	35,564	1,663	33,643
			(16)	(36,121)	(59,243)	(9,119)	(58,701)
2002	10 per cent	1,112,807	41	25,641	34,793	1,595	32,898
			(16)	(32,321)	(194,785)	(8,173)	(194,587)
	Population	11,131,065	41	25,650	34,583	1,598	32,679
			(16)	(32,085)	(79,864)	(9,460)	(81,747)
2001	10 per cent	1,095,858	41	25,028	34,074	1,469	32,283
			(16)	(30,159)	(271,918)	(9,155)	(271,454)
	Population	10,960,782	41	25,055	33,854	1,468	32,050
			(16)	(34,047)	(110,136)	(15,253)	(109,623)
2000	10 per cent	1,076,253	41	23,889	32,630	1,320	31,031
			(16)	(28,676)	(217,797)	(8,081)	(217,517)
	Population	10,767,219	41	23,879	32,402	1,314	30,864
			(16)	(29,745)	(88,956)	(7,873)	(215,040)
1999	10 per cent	1,056,573	41	22,905	31,026	1,576	29,654
			(16)	(26,995)	(45,294)	(13,732)	(117,083)
	Population	10,569,005	41	22,893	30,948	1,595	29,413
			(16)	(28,541)	(44,110)	(19,339)	(59,025)
1998	10 per cent	1,048,282	41	21,954	29,733	1,153	28,495
			(16)	(24,728)	(38,422)	(11,665)	(38,241)
	Population	10,483,382	41	21,945	29,703	1,177	28,446
			(16)	(25,112)	(39,605)	(20,131)	(42,637)
1997	10 per cent	1,045,594	41	20,836	28,254	1,412	26,839
			(16)	(23,035)	(34,405)	(12,120)	(34,798)
	Population	10,452,630	41	20,844	28,237	1,446	26,789
			(16)	(23,383)	(45,092)	(20,893)	(40,929)
1996	10 per cent	1,034,426	40	20,054	26,911	1,412	25,549
			(16)	(22,064)	(41,132)	(12,380)	(41,390)
	Population	10,345,854	40	20,035	26,826	1,438	25,456
			(16)	(21,942)	(35,368)	(20,474)	(41,775)
1995	10 per cent	1,012,617	40	18,928	25,329	1,210	24,212
			(16)	(20,605)	(29,223)	(11,073)	(29,898)
	Population	10,126,599	40	18,920	25,324	1,236	24,182
			(16)	(20,739)	(30,769)	(19,568)	(35,442)
1994	10 per cent	989,873	40	18,112	24,346	1,179	23,290
			(16)	(19,640)	(28,203)	(11,188)	(29,158)
	Population	9,897,579	40	18,107	24,324	1,206	23,233
			(16)	(19,904)	(40,131)	(18,210)	(42,041)
1993	10 per cent	977,560	40	17,512	23,319	1,170	22,206
			(16)	(18,543)	(25,462)	(11,931)	(26,896)
	Population	9,778,261	40	17,505	23,293	1,226	22,129
			(16)	(18,657)	(47,152)	(88,086)	(100,265)
1992	10 per cent	979,061	40	17,171	22,606	1,341	21,294
			(16)	(18,008)	(24,667)	(10,822)	(26,057)
	Population	9,792,088	40	17,169	22,590	1,370	21,262
			(16)	(18,613)	(35,136)	(15,508)	(48,438)
1991	10 per cent	983,470	40	16,935	22,247	702	21,733
			(16)	(17,537)	(24,661)	(10,510)	(25,181)
	Population	9,840,457	40	16,928	22,229	718	21,705
			(16)	(17,831)	(26,873)	(13,477)	(28,434)

Source: Author's calculations. Note: Standard deviations in parentheses.