


# NATSEM

National Centre for Social and Economic Modelling

• University of Canberra •



## **APPSIM – Selection of the Main Source Data File for the Base Data**

**Simon Kelly**

**Working Paper No. 2**

**April 2007**



## **About NATSEM**

The National Centre for Social and Economic Modelling was established on 1 January 1993, and supports its activities through research grants, commissioned research and longer term contracts for model maintenance and development with the federal departments of Family and Community Services, Employment and Workplace Relations, Treasury, and Education, Science and Training.

NATSEM aims to be a key contributor to social and economic policy debate and analysis by developing models of the highest quality, undertaking independent and impartial research, and supplying valued consultancy services.

Policy changes often have to be made without sufficient information about either the current environment or the consequences of change. NATSEM specialises in analysing data and producing models so that decision makers have the best possible quantitative information on which to base their decisions.

NATSEM has an international reputation as a centre of excellence for analysing microdata and constructing microsimulation models. Such data and models commence with the records of real (but unidentifiable) Australians. Analysis typically begins by looking at either the characteristics or the impact of a policy change on an individual household, building up to the bigger picture by looking at many individual cases through the use of large datasets.

It must be emphasised that NATSEM does not have views on policy. All opinions are the authors' own and are not necessarily shared by NATSEM.

Director: Ann Harding

ISSN 1834-7630  
ISBN 978-1-74088-266-8

© NATSEM, University of Canberra 2007

National Centre for Social and Economic Modelling  
University of Canberra ACT 2601  
Australia

170 Haydon Drive  
Bruce ACT 2617

Phone + 61 2 6201 2780  
Fax + 61 2 6201 2751

Email [natsem@natsem.canberra.edu.au](mailto:natsem@natsem.canberra.edu.au)

Website [www.natsem.canberra.edu.au](http://www.natsem.canberra.edu.au)

Title *APPSIM – Selection of the Main Source Data File for the Base Data*  
Author Simon Kelly  
Series Working Paper No. 2



## **APPSIM Working Paper Series**

The Australian Government has identified that future government outlays will exceed future government revenues. Resolving this budget shortfall will require either higher taxes upon future generations or reductions in spending programs (or some combination of these). The Australian Government currently has only limited ability to assess the future distributional and revenue consequences of changes in tax and outlay programs.

NATSEM in collaboration with 13 Government organisations and two international academics is developing a new dynamic microsimulation model, APPSIM, to enhance the planning and policy simulation capacity of the Australian Government.

The APPSIM working paper series comprises papers covering the construction of APPSIM.

### **Author note**

Simon Kelly is an Associate Professor at the University of Canberra and a Principal Research Fellow at the National Centre for Social and Economic Modelling.

### **Acknowledgments**

The author would like to gratefully acknowledge the funding provided by the Australian Research Council (under grant LP0562493), and by the 13 research partners to the grant : Treasury; Communications, Information Technology and the Arts; Employment and Workplace Relations; Health and Ageing; Education, Science and Training; Finance and Administration; Families, Community Services and Indigenous Affairs; Industry, Tourism and Resources; Immigration and Multicultural Affairs; Prime Minister and Cabinet; the Productivity Commission; Centrelink; and the Australian Bureau of Statistics. The author would like to acknowledge and thank all the members of the APPSIM Technical Advisory Group for guidance and helpful comments provided on earlier draft versions of the paper.

We would also like to acknowledge our two partner investigators, Professors Jane Falkingham and Maria Evandrou of the University of Southampton.

# Contents

<b>APPSIM Working Paper Series</b>	<b>iii</b>
<b>Author note</b>	<b>iii</b>
<b>Acknowledgments</b>	<b>iii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 The options</b>	<b>2</b>
2.1 2001 Census	2
2.2 2003-04 HES & SIH	3
2.3 HILDA	3
<b>3 Comparison</b>	<b>4</b>
<b>4 Recommendation</b>	<b>6</b>
<b>References</b>	<b>7</b>
<b>A 2001 Census HSF Variables</b>	<b>8</b>

# 1 Introduction

Microsimulation models are usually based on large random samples of a population of individuals, households or firms. These initial random samples are referred to as the base data. In most dynamic microsimulation models, such as APPSIM, the model starts with a population of individuals constructed from empirical data with attributes describing their state at the starting date (Zaidi and Scott 2001). The simulation then ‘ages’ or moves them forward through time, updating individuals’ attributes in accordance with defined transition or event processes (for example the probability that a woman of a certain age will give birth to a child) (Harding 1993). After every unit has been aged and the transitions applied, the process is repeated, thus advancing the samples through simulated time.

Given that the base data provide the starting point for the simulation, the selection of the source of the base data is a key defining element of APPSIM. There are a number of requirements for the base data. The first requirement is that the base data must be available to researchers in a unit record format which can be manipulated as part of a complex model – that is, records must be available for separate individuals, whose characteristics can be changed as the model progresses. Second, the data must have adequate coverage and be representative of the Australian population. As representativeness is related to sample size, this second requirement effectively says that the base data must be from a large sample. An adjunct to this requirement is that the data must be accepted by users as being representative. The third requirement of the base data is that as many as possible of the attributes required for the simulation are provided in the base data and are specified in sufficient detail for the simulation. The final requirement is that ideally the base data need to contain a certain amount of historical information. One of the strengths of dynamic microsimulation models is that they are able to utilise duration, current attributes and historical information in estimating future behaviour. It is important that estimates calculated at the start of the simulation be able to access the full breadth of attributes.

Typically, these requirements cannot be fulfilled from a single data source – and this will also be the case with APPSIM. To overcome the shortcomings of a selected data source, other data sources are used to supplement the main data by providing extra attributes or expanding the detail of an attribute already available. This enhanced data file will then become the base data for APPSIM.

The purpose of this note is to describe the options for this main source of the base data and to provide a recommendation on the main source of the base data for APPSIM.

## 2 The options

A study of surveys and censuses in Australia shows that there is no single data source in Australia that meets all of the requirements listed above – representative, detailed and containing historical attributes. However, there are three potential main sources – two from the Australian Bureau of Statistics (ABS) and the third from the Department of Families, Community Services and Indigenous Affairs (FaCSIA). The files are:

1. the 2001 Census of Population and Housing Household Sample File (HSF);
2. the 2003-04 Household Expenditure Survey (HES) and Survey of Income and Housing (SIH) Confidentialised Unit Record Files; and
3. the Household, Income and Labour Dynamics (HILDA) survey.

### 2.1 2001 Census

The 2001 Census of Population and Housing HSF is a unit record file of a representative one per cent sample derived from the 2001 Census (ABS 2003). Some of the features of the data file are:

- the 2001 Census HSF is a 1% sample of private dwellings, with their associated family and person records, and a 1% sample of persons from all non-private dwellings together with a record for the non-private dwelling;
- there are records for 188,013 persons, 75,451 dwellings and 79,320 families;
- the weights for each record are constant – 100 for each person record and 4.07 for each dwelling record;
- almost complete population coverage is provided;
- the Census has a far lower non-response rate than other surveys;
- some attributes are collected and recorded in less detail in the Census than they are in other surveys. For example, weekly household income is assigned one of 15 categories (such as '\$200-299') and the age for a sample person record is stored in one of 38 categories (with the highest being '85 years and over');
- on the other hand, the Census includes more detail on aspects such as linkage within the family; and
- some very limited historical data are recorded (such as, where the person lived one and five years before the Census).

## 2.2 2003-04 HES & SIH

Prior to 2003-04, the HES and SIH were conducted independently (ABS 2006). The HES sample was drawn from dwellings not recently included in an ABS household survey, whereas the SIH sample was drawn from dwellings that had just completed eight months participation in the Monthly Population Survey. In 2003-04, the HES was integrated with the SIH. The integration of the two surveys has lowered respondent burden and the resultant dataset is richer because HES and SIH results are more comparable than previously. The features of these files are:

- the sampling for the files excluded non-private dwellings (such as hospitals, institutions, nursing homes, hotels, hostels, etc.) and dwellings in areas defined as very remote or Indigenous communities;
- the SIH contains responds from 22,286 persons and 11,361 households with information on personal and household characteristics, detailed income by source, and detailed information on assets and liabilities: a subset of 13,726 persons and 6,957 households were further selected in the HES and asked to supply detailed information on household expenditure, loans and financial stress;
- for the basic CURF (the CD-ROM version of the source data), persons were removed from all households with 7 or more persons to reduce them to a maximum household size of 6. This also resulted in the deletion of several whole income units, mainly comprising a single person record only. A total of 88 persons aged under 15 years were dropped and 29 persons aged 15 years or over were dropped;
- the weights are different for each record on the SIH and HES;
- due to the sample size, the detail of some attributes has been reduced. For example, for the variable 'State of usual residence', the ACT and the NT have been combined as ACT/NT. All income items, some expenditure items relating to housing, and some loan data have been perturbed, while other variables have had values ranged, collapsed or top-coded;
- the HES and SIH, as would be expected, include far richer (more detailed and comprehensive) data on people's labour force characteristics, incomes, expenditure, assets and liabilities; and
- the HES and SIH link individuals within households, families and income units.

## 2.3 HILDA

The Household, Income and Labour Dynamics in Australia (HILDA) Survey is a household-based panel study which began in 2001 (Melbourne Institute 2006). The

HILDA Survey was initiated, and is funded, by the Australian Government through FaCSIA. Responsibility for the design and management of the survey rests with a group led by the Melbourne Institute of Applied Economic and Social Research. Data collection has been sub-contracted to ACNielsen, a private market research company.

HILDA has the following key features:

- it collects information about economic and subjective well-being, labour market dynamics and family dynamics;
- the sampling for the initial wave excluded those in non-private dwellings and dwellings in very remote areas;
- the wave 1 panel consisted of 7,682 households and 19,914 individuals. The panel members are followed over time providing longitudinal data;
- special questionnaire modules are included each wave;
- the funding has been guaranteed for the first eight waves;
- the weights are different for each record on HILDA and vary between waves;
- missing data is often imputed;
- the survey does not have the same official status as ABS surveys;
- aggregated data does not necessarily match up with externally sourced aggregates; and
- some particularly useful data are only collected in occasional 'special modules' (for example wealth and liability data are only being collected in wave 2 and wave 6). This reduces the usefulness of the longitudinal aspects in some cases.

### **3 Comparison**

All three sources have important strengths and weaknesses as base data for APPSIM. The Census provides the best coverage but has the lowest level of detail for some important attributes. The HES and SIH surveys provide the richest detail in regard to income and expenditure but the weights vary between records and only private dwellings are covered. The HILDA data has the broadest range of current and historical attributes but has the smallest sample size, only covers private dwellings and is not as well established as the ABS surveys.

NATSEM has previously constructed two dynamic microsimulation models – the HARDING dynamic cohort microsimulation model (Harding 1993) and the DYNAMOD dynamic population microsimulation model (King et al. 1999; Kelly and

King 2001). This previous experience has led us to the conclusion that it is particularly important to have a very large base data file. There are two key reasons for this. The first, as Zaidi and Scott explain, is that “sample size is related to representativeness, in that proportionately small subgroups of the population cannot be adequately represented in small samples. Thus, the base dataset must have a large enough sample size to allow reliable disaggregation of simulations across subgroups of interest” (2001, p. 5). Our experience with analyses of cross-sectional sample surveys of 10,000 households or less suggests that such a sample size does not permit adequate analysis of relatively rare population sub-groups (such as large sole parent families).

The second is that larger base data sets improve the accuracy of the simulation processes. Dynamic models typically rely on Monte Carlo simulation to select the individuals to whom changes in state will occur. For example, suppose that an econometric estimation process has indicated that one per cent of a particular population sub-group should be selected to shift from being ‘unemployed’ to being ‘employed’. Typically, this is achieved by comparing this probability against a randomly generated uniformly distributed number, where those in the sub-group whose random number is equal to or less than 0.01 are chosen to actually make the transition. Other factors being equal, a larger sample size means that more records are actually selected to make the change, reducing the degree of Monte Carlo error and allowing more finely-grained simulation of life events.

On balance, therefore, the Census HSF appears to be the best solution. While not meeting all of the requirements in a single source of base data, the perceived advantages of the Census HSF are that it provides the greatest coverage of the Australian population (it has the best coverage of older Australians, covers all areas of Australia and covers both private and non-private dwellings); it has the largest sample size (almost ten times the other surveys); and it uses a constant weight for each record. This constant weight makes simulation of the marriage market considerably easier as each record can be matched to another single record. Records with differing weights are considerably harder to match together.

As noted above, the Census has a number of limitations but a number of these can be overcome by imputing data from other surveys. These surveys can be used to provide more detail to some of the base data attributes and to impute more information onto the base data.

## **4 Recommendation**

Our decision is to use the 2001 Census of Population and Housing HSF as the main source of data for the base file of APPSIM. Attachment A shows the variables available in the 2001 one-per cent Census CD-ROM file.

## References

- Australian Bureau of Statistics 2003, *Census of Population and Housing Household Sample File Australia 2001*, Technical Paper, Cat No. 2037.0, Australian Bureau of Statistics, Canberra, September.
- Australian Bureau of Statistics 2006, *Household Expenditure Survey and Survey of Income and Housing - Confidentialised Unit Record Files*, Technical Paper Australia 2003–04, Cat No. 6540.0.00.001, Australian Bureau of Statistics, Canberra, June.
- Harding, A. 1993, *Lifetime Income Distribution and Redistribution. Applications of a Microsimulation model*, Contributions to Economic Analysis, North-Holland, Elsevier Science Publishers, the Netherlands.
- Kelly, S. and King, A. 2001, 'Australians over the coming 50 years: providing useful projections', *Brazilian Electronic Journal of Economics*, vol. 4, no. 2, pp. 1-23.
- King, A., Bækgaard, H. and Robinson M. 1999, *The Base Data for DYNAMOD-2*, Technical Paper No. 20, National Centre for Social and Economic Modelling, University of Canberra, Canberra, December.
- Melbourne Institute 2006, *HILDA User Manual – Release 4*, Melbourne Institute of Applied Economic and Social Research, University of Melbourne, March.
- Zaidi, A. and Scott, A. 2001, *Base Dataset for the SAGE Model*, SAGE Technical Note No. 1, ESRC SAGE Research Group, London School of Economics, United Kingdom, September.

## A 2001 Census HSF Variables

### *Person Characteristics*

Age (AGEP)  
Ancestry (ANCP)  
Australian Citizenship (CITP)  
Birthplace of Female Parent (BPFP)  
Birthplace of Individual (BPLP)  
Birthplace of Male Parent (BPMP)  
CD of Usual Residence Census Night (CDUCP)  
Child Type (CTPP)  
Computer Use at Home (COMP)  
Family/Household Reference Person Indicator (RPIP)  
Full/Part-Time Student Status (STUP)  
Highest Level of Schooling Completed (HSCP)  
Hours Worked (HRSP)  
Indigenous Status (INGP)  
Individual Income (weekly) (INCP)  
Industry of Employment (INDP)  
Industry Sector (GNGP)  
Internet Use (NETP)  
Journey to Work: Destination Zone (JTWDZNP)  
Journey to Work: Study Area (JTWSAP)  
Labour Force Status/Status in Employment (LFSP)  
Language Spoken at Home (LANP)  
Method of Travel to Work (MTWP)  
Non-School Qualification: Field of Study (QALFP)  
Non-School Qualification: Level of Education (QALLP)  
Non-School Qualification: Year Completed (QALYP)  
Occupation (OCCP)  
Postal Area of Usual Address Census Night (POCUCP)  
Proficiency in Spoken English (ENGP)  
Proficiency in Spoken English/Language (ENGP01)  
Registered Marital Status (MSTP)  
Relationship in Household (RLHP)  
Religious Affiliation (RELP)  
Residential Status in Non-Private Dwelling (RLNP)  
Sex (SEXP)  
SLA of Usual Residence Census Night (SLAUCP)  
SLA of Usual Residence Five Years Ago (SLAU5P)  
SLA of Usual Residence One Year Ago (SLAU1P)  
Social Marital Status (MDCP)  
State of Usual Residence Census Night (STEUCP)  
State of Usual Residence Five Years Ago (STEU5P)  
State of Usual Residence One Year Ago (STEU1P)  
Type of Educational Institution Attending (TYPP)

Usual Address Five Years Ago Indicator (UAI5P)  
Usual Address Indicator Census Night (UAICP)  
Usual Address One Year Ago Indicator (UAI1P)  
Year of Arrival in Australia (YARP)

### *Household/Dwelling Characteristics*

Count of Persons Temporarily Absent from Household (CPAD)  
Dwelling Location (DLOD)  
Dwelling Structure (STRD)  
Dwelling Type (DWTD)  
Household Five Year Mobility Indicator (MV5D)  
Household Income (weekly) (HIND)  
Household Income Derivation Indicator (HIDD)  
Household One Year Mobility Indicator (MV1D)  
Household Type (HHTD)  
Housing Loan Repayments (monthly) (HLRD) dollar values  
Housing Loan Repayments (monthly) (HLRD01) ranges  
Landlord Type (LLDD)  
Number of Bedrooms in Private Dwellings (BEDD)  
Number of Motor Vehicles (VEHD)  
Number of Motorbikes and Scooters (MCYCD)  
Rent (weekly) (RNTD) dollar values  
Rent (weekly) (RNTD01) ranges  
Tenure Type (TEND)  
Type of Non-Private Dwelling (NPDD)

### *Family Characteristics*

Count of Dependent Children Under 15 Temporarily Absent (CDCAF)  
Count of Dependent Students (15–24) Temporarily Absent (CDSAF)  
Count of Non-Dependent Children Temporarily Absent (CNDAF)  
Count of Persons Temporarily Absent from Family (CPAF)  
Family Characteristics  
Family Income (weekly) (FINF)  
Family Income Derivation Indicator (FIDF)  
Family Number (FNOF)  
Family Type (FMTF)  
Location of Spouse (SPLF)