CONFIDENTIAL

New techniques for household microsimulation, and their application to Australia

John Richard Cumpston

14 December 2011

A thesis submitted for the degree of Doctor of Philosophy of the Australian National University

STATEMENT ON ORIGINAL WORK

I declare that this thesis is an original work, and is an account of research carried out by myself while enrolled as a PhD candidate at The Australian National University. This thesis does not contain material that has been accepted for the award of any other degree or diploma in any University, nor material published or written by another person, except where due reference is made.

John Richard Cumpston

ACKNOWLEDGEMENTS

I am grateful for the help given to me by

- David Service, who chaired my supervisory panel, and Roger Bradbury, Tim Higgins and Jeromey Temple, who were members of the panel at various times.

- The International Microsimulation Association, whose conferences in Vienna in 2007, Ottawa in 2009 and Stockholm in 2011 provided valuable opportunities to learn about household microsimulation models and meet other modellers.

- Marcia Keegan, Simon Kelly and Richard Percival of APPSIM, Marcello Morciano of CAPP_DYN, John Sabelhaus and Michael Simpson of CBOLT, Richard Morrison of DYNACAN, Seiichi Inagaki of INAHSIM, Eric Miller of ILUTE, Gijs Dekkers of MIDAS-Belgium, Trude Gunnes and Nils Stolen of MOSART, Mark Birkin and Belinda Wu of Moses, Howard Redway of Pensim, Tomas Pettersson of SESIM and Einar Holm and Kalle Makila of SVERIGE, for information about the microsimulation models they have helped create.

- Richard Easther, Cathal O'Donoghue, John O'Leary and Paul Williamson for information about microsimulation techniques.

- Reviewers for the International Journal of Microsimulation, who suggested many improvements to papers submitted for publication.

- Gijs Dekkers, whose draft paper on weights in dynamic-ageing microsimulation models led to the simulation tests described in chapter 5, and to a joint presentation at the 2011 International Microsimulation Association Conference. Comments in sections 5.5 and 5.6 relate to test results for the method proposed by Dekkers, and are largely taken from a joint paper with him.

.

- Hugh Sarjeant, who co-authored papers with me on fertility, mortality and regional migration, and who wrote a program creating synthetic households from Australian census data for statistical local areas.

- The International Institute for Applied Systems Analysis in Vienna, who gave me a research position for seven weeks, allowing me to study their demographic, energy and agriculture models.

- Paul Thomson and Corey Plover of Cumpston Sarjeant Pty Ltd, who advised me about C# programming techniques and data access methods.

- Phillip Gallagher and Anthony King, who helped me understand the Australian Treasury's potential uses for household microsimulation.

- The Department of Families, Housing, Community Services and Indigenous Affairs, and the Melbourne Institute of Applied Economic and Social Research, for the use of the HILDA dataset and the assistance provided in using the dataset.

ABSTRACT

Household microsimulation models are sometimes used by national governments to make long-term projections of proposed policy changes. They are costly to develop and maintain, and sometimes have short lifetimes. Most present national models have limited interactions between agents, few regions and long simulation cycles. Some models are very slow to run. Overcoming these limitations may open up a much wider range of government, business and individual uses.

This thesis suggests techniques to help make multi-purpose dynamic microsimulations of households, with fine spatial resolutions, high sampling densities and short simulation cycles. Techniques suggested are

- simulation by sampling with loaded probabilities
- proportional event alignment
- event alignment using random sampling
- immediate matching by probability-weighting
- immediate “best of n” matching.

All off these techniques are tested in artificial situations. Three of them - sampling with loaded probabilities, alignment using random sampling and best of n matching - are successfully tested in the Cumpston model, a household microsimulation model developed for this thesis.

Sampling with loaded probabilities is found to give almost identical results to the traditional all-case sampling, but be quicker. The suggested alignment and matching techniques are shown to give less distortion and generally lower runtimes than some techniques currently in use.

The Cumpston model is based on a 1% sample from the 2001 Australian census. Individuals, families, households and dwellings are included. Immigration and emigration are separately simulated, together with internal migration between 57 statistical divisions. Transitions between 8 person types are simulated, and between 9 occupations. The model projects education, employment, earnings and retirement savings for each individual, and dwelling values, rents and housing loans for each household. The onset and development of diseases for each individual are simulated,

Validation of the model was based on methods used by the Orcutt, CORSIM, DYNACAN and APPSIM models. Iterative methods for model calibration are described, together with a statistical test for creep in multiple runs.

The model takes about 85 seconds to make projections for 50 years with yearly simulation cycles. After standardizing for sample size and projection years, this is a little slower than the fastest national models currently operating.

A planned extension of the model is to 2.2 million persons over 2,214 areas, synthesized from 2011 census tabulations. Using multithreading where feasible, a 50-year projection may take about 10 minutes.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Scope of this thesis

This thesis notes the long history and potentially wide uses of household microsimulation models. Limited capabilities and slow runtimes are restricting present uses. New simulation methods are suggested, intended to reduce runtimes and distortion, and to allow shorter simulation cycles:

- simulation by sampling with loaded probabilities
- proportional event alignment
- event alignment using random sampling
- immediate matching by probability-weighting
- immediate “best of n” matching.

All off these techniques are tested in artificial situations. Three of them - sampling with loaded probabilities, alignment using random sampling and best of n matching - are successfully tested in the Cumpston model, a household microsimulation model developed for this thesis.

1.2 Orcutt proposals for microsimulation, and their implementation

Orcutt (1957) noted the limitations of aggregate models of socio-economic systems, and proposed the modelling of “actual decision making-units of the real world, such as the individual, the household and the firm”. Their behavior would be recursively simulated over short periods, such as a week or a month, thus avoiding the need to solve any simultaneous equations. Incorporating a much wider range of behavior was expected to assist prediction-making in many ways.

Orcutt, Greenberger, Korbel & Rivlin (1961) described a model based on Orcutt’s proposals, but modelling only individuals and households. Although restricted by limited computer capabilities, this model introduced seven techniques which have since been widely used in household microsimulations (section 2.9). Like some subsequent models, the practical value of the model was reduced by a serious programming error.

1.3 Limited range of uses of present models

Although household microsimulation models have been built for many countries, their capabilities and uses are generally limited. Simulation intervals are usually years, rather than the shorter periods considered desirable by Orcutt. Although Caldwell (1983 p61) noted the policy analysis payoffs from fine spatial disaggregation, most models have few regions. Models are largely used to explore the long-term consequences of changes to tax and social security systems. Long simulation intervals, slow run times, limited spatial disaggregation, uncertain reliability and high costs may be preventing a much broader ranges of uses.

1.4 Using sampling with loaded probabilities to get shorter simulation intervals

Household models have almost always simulated the occurrence of each type of event for each person each simulation interval. This results in run times increasing almost linearly with the number of intervals per year. This thesis suggests the use of simulation by sampling with loaded probabilities, a form of stratified sampling. This technique greatly reduces the run time penalties from short simulation intervals, allowing intervals as short as a day. Short intervals are likely to be needed in modelling agent interactions, for example in dwelling sales.

1.5 Ways to reduce run times

After adjusting for sample size and projection years, there are large differences between the run times of current household models (section 2.14). Some of this variation reflects different functionality. The choice of programming language, and the use of object-oriented programming, does not appear to be responsible for much of the variation. The slowest run times appear to result from over-complex state alignment. Modellers may need to resist requests for close alignment with independently derived state totals, as this may result in distortions as well as very slow run times. Multi-threading was used to a limited extent by McKay (2003 p3), and more extensively by Fredricksen, Knudsen and Stolen (2011). This thesis suggests sampling, event alignment and matching techniques which may substantially reduce run times.

1.6 Ways to increase spatial disaggregation

This thesis suggests the use of logistic regression models for movements between large regions, with matching of households to individual dwellings within regions. Each spatial region may require at least 1,000 persons to reasonably represent it, so that larger numbers of regions require larger sample sizes, and low runtimes per person.

1.7 Ways to reduce distortions from event alignment and matching

Some distortions occur because of inherently faulty event alignment procedures. This thesis suggests a modification to the widely used two-pass procedure of Johnson (2001), intended to slightly reduce its distortion. The thesis also suggests alignment using random selection, a one-pass procedure which should give similar results to the modified form of Johnson's procedure, and be suitable with short simulation intervals.

Matching procedures are needed for many different types of household change, including partnership formations. Procedures are also needed to match moving households with available dwellings. The thesis suggests two new forms of immediate matching - probability-weighted selection from all available matches, and "best of n" selection. A range of statistical measures and graphs are used to show that both of these techniques are likely to give better results than commonly used batch matching procedures. Best of n matching is a particularly quick matching procedure.

1.8 Programming languages and data structures

Many different programming languages have been used for household microsimulation models (section 2.14). Percival (2007) recommended C# for use by APPSIM, rather than the faster C++, because the design of C# helps develop code that is safer and easier to maintain. C# was used for the Cumpston model, and its multithreading capabilities may be important in future extensions to the model.

Initial trials with a database structure gave unacceptably slow simulations. Instead, persons and dwellings are stored in indexed arrays, intended to avoid duplicated storage and allow rapid access to persons and dwellings with particular characteristics. For example, rapid access can be obtained to all the persons in a state, age-band and household type, and to all vacant or occupied dwellings in a region. These index structures are designed to meet the needs of the random access sampling, alignment and matching procedures developed in the thesis. Although requiring some initial programming, these indexed arrays have proved simple and robust in practice.

1.9 Ways to reduce errors

A serious error occurred in the model described by Orcutt et al (1961 p203), and large errors have occurred in more recent models, including ILUTE (Miller et al 2008 p28). The Orcutt et al model provided optional listings of the details of each simulated event, and this error-detection technique has been used by many subsequent models. Morrison (2008) described the wide range of techniques used to validate DYNACAN, and many of these techniques were applied to the Cumpston model, including putting all assumptions in a well-annotated input file. A statistical test was implemented to detect creep in multiple runs.

1.10 Data needs and sources

Data are needed for baseline populations, transition assumptions and temporal changes. Baseline populations are here derived from a 1% sample of the Australian 2001 census (ABS 2003d), supplemented by the 2000-01 Survey of Income and Housing Costs (ABS 2003b), the 2003 Survey of Disability, Ageing and Carers (ABS 2005), and the HILDA Longitudinal Survey (Watson 2008). HILDA was also a very useful source of transition assumptions, supplemented for some purposes with detailed administrative data.

1.11 Simulating demographic events and household changes

Fertility, mortality, emigration and immigration assumptions for the Cumpston model are chosen so as to replicate ABS 50-year population projections (ABS 2003c). Characteristics of females who have recently given birth, and of persons who have recently immigrated to Australia, are obtained by analysis of the unit records for the 1% sample of the 2001 census. Probabilities of persons moving between statistical divisions are obtained by logistic analysis of data on movements between the 1991 and 1996 censuses, and between the 1996 and 2001 censuses.

Assumptions about household changes, and about households moving within statistical divisions, are derived by comparing successive waves of the HILDA longitudinal survey. Non-responses, survey losses and limited sample sizes make it very difficult to derive reliable estimates for all the necessary probabilities, and some

adjustments are made to obtain reasonable balances with ABS household projections (ABS 2004).

1.12 Simulating education, occupation, employment and earnings

Entry, progression and exit assumptions for schools are chosen so as to reproduce ABS data on school enrolments (ABS 2011f). Entry, completion and dropout rates for technical and further education are estimated from historical data on entries, enrolments and qualifications (NCVER 2011), and for university education from similar data (DEEWR 2011a & b). Assumptions about movements in and out of the labour force, between employment and unemployment, and between occupations, are based on HILDA data. Models of earnings from employment and self-employment, and of year by year changes in these earnings, are also based on HILDA data.

1.13 Simulating superannuation contributions and fund balances

Under Australian legislation, superannuation funds are intended to provide retirement benefits separately from social security entitlements. Superannuation assumptions are here based on administrative data from APRA (the Australian Prudential Regulatory Authority) and ATO (the Australian Taxation Office), and on the responses to the superannuation questions in HILDA waves 2 and 6. Major legislative changes have occurred in the last 5 years, and the contribution and befit models were iteratively adjusted to approximately match administrative totals in the 9 years to 30/6/10.

1.14 Simulating housing

The values of owner-occupied dwellings in the 2001 census sample are estimated from a regression relationship fitted to 2000-01 SIHC data, and the values of rented dwellings in the census sample are estimated from REIA (2008) gross rental yields. Dwellings are modelled separately from households, with construction and demolition rates based on the vacancy rate for each statistical division. New and moving households are matched to vacant dwellings, using best of n matching.

1.15 Simulating disease, disability and need for residential care

Incidence, development and mortality are simulated for 123 diseases, using assumptions based on AIHW disease models (Begg et al 2007). These simulations provide estimates of persons with severe or profound activity limitations, and of persons likely to need residential care. Comparisons with numbers reporting each disease, dying from each disease, and of entry rates to residential care, give mixed results. Use of standardized disease classifications by ABS, AIHW and the Department of Health and Ageing would be helpful.

1.16 Meeting the needs of potential users

If fast models with many regions and short simulation cycles can be provided, potential users include state and local governments, financial institutions and other businesses, churches, non-government organizations and individuals. Although some initial government funding may be needed, independent research institutes and consultancies may be better long-term hosts.

1.17 Assistance to other modellers

The developers of models for Japan and Toronto have expressed interest in using the immediate matching techniques described in this thesis (Inagaki 2009c, Miller 2009b). The Cumpston model was adapted to test a proposal by Dekkers for the use of weighted records. and a paper about this has been presented (Dekkers & Cumpston 2011). A paper on alignment and matching techniques has been published by the International Microsimulation Journal (Cumpston 2010a), and a paper on sampling with loaded probabilities is under consideration by that journal. A paper on disease and long-term care microsimulation has been presented (Cumpston 2011), and accepted for publication in a book on applications of microsimulation.

2. LITERATURE REVIEW

This chapter defines key terms, and reviews relevant literature. Emphasis is given to the initial emergence of new concepts, and on the limitations of current models.

Details are given for 26 national cross-sectional dynamic closed household models. There have been historic links between many of these models, with many common methods. Many of the models have been costly to develop, and some of them are no longer operational. They are largely being used for long-term national purposes.

Nearly all the household microsimulation techniques in use today were developed by Orcutt, Greenberger, Korbel & Rivlin (1961), or by researchers working in the 1950s and 1960s. Current models lack the full range of agents proposed by Orcutt (1957).

For 15 recent models with run times available, there is a range of about 3,600 to 1 in standardized run times. Some of these variations are due to different functionalities, computer speeds and computer languages, but some slow run times may be due to data access procedures or excessive alignment.

Most present national models have limited interactions between agents, long simulation cycles and few regions. Overcoming these limitations may open up a much wider range of government, business and individual uses.

2.1 Definitions of microsimulation

The International Microsimulation Association (2009) defines microsimulation as

> *"…a modelling technique that operates at the level of individual units such as persons, households, vehicles or firms. Within the model each unit is represented by a record containing a unique identifier and a set of associated attributes - e.g. a list of persons with known age, sex, marital and employment status; or a list of vehicles with known origins, destination and operational characteristics. A set of rules (transition probabilities) are then applied to these units leading to simulated changes in state and behavior."*

Statistics Canada (2009) provides a similar definition

> *"Microsimulation models are computer models that operate at the level of the individual behavioural entity, such as a person, family, or firm. Such models simulate large representative populations of these low-level entities in order to draw conclusions that apply to higher levels of aggregation such as an entire country."*

2.2 Orcutt's 1957 proposals for household and firm simulations

Microsimulations of United States households and firms were proposed by Orcutt (1957) and implemented for 4,580 households by Orcutt, Greenberger, Korbel & Rivlin (1961). They commented (p6):

> *"Our socio-economic system is a complicated structure containing millions of interacting units, such as individuals, households and firms. It is these units which actually make decisions about spending and saving, investing and producing, marrying and having children. It seems reasonable to expect that*

our predictions would be more successful if they were based on knowledge about these elemental decision-making units - how they behave, how they respond to changes in their circumstances, and how they interact."

Although proposing a “stochastic microanalytic model of a socioeconomic system” including individuals, families, firms, banks, labour unions and local governments, they implemented only a household model.

The microsimulations proposed by Orcutt were examples of Monte Carlo simulations, using random numbers to simulate event occurrences. The term “Monte Carlo method” was suggested in late 1946 or early 1947 by Metropolis, in relation to the modelling of neutron diffusion in fissionable material, using the ENIAC computer. Similar calculations had been done by Fermi in the early thirties, when studying the moderation of neutrons (Metropolis 1987).

2.3 Some national cross-sectional dynamic models of households

Table 2.1 lists 26 national models, including some of historic importance (Orcutt, DYNASIM, CORSIM, DYNACAN).

Table 2.1 National cross-sectional dynamic household models

Model name	Country	Status	First version	Sampling density	Million persons	Number regions
APPSIM	Australia	Developing	2009	0.9%	0.18	1
CAPP-DYN	Italy	Active	2004?	0.3%	0.17	1
CBOLT	USA	Active	2002?	0.1%	0.30	1
CORSIM	USA	Inactive	1988?	0.07%	0.18	1
Cumpston	Australia	Developing	2009	0.9%	0.175	57
Destinie	France	Active	1995?	0.1%	0.065	
DYNACAN	Canada	Inactive	1998?	0.9%	0.213	2
DYNAMOD2	Australia	Inactive	2002?	0.9%	0.15	1
DYNASIM	USA	Active	1975			
INAHSIM	Japan	Active	1986	0.1%	0.128	1
MASS	USA	Inactive	1979?			
MICROHUS	Sweden	Inactive	1989?		0.005	
MIDAS-BE	Belgium	Active	2008	3%	2.0	1
MOSART	Norway	Active	1990	100%	7.2	1
NEDYMAS	Netherlands	Inactive	1990?		0.01	
Moses	UK	Developing	2008?	100%	0.75	33
Orcutt	USA	Inactive	1959	0.01%	0.01	1
PENSIM	USA	Inactive				
Pensim2	UK	Active	2004	0.1%	0.06	1
POLISIM	USA	Active	2001?	0.1%	0.179	
PRISME	France	Active	2006?	5%	3.2	
SAGE	UK	Inactive	2003	0.1%	0.054	1
SESIM	Sweden	Active	1997	0.12%	0.112	9
Sfb3	FDR	Inactive	1974		0.63	
SMILE	Ireland	Active	2005?	50%	2.1	30
SVERIGE	Sweden	Inactive	2002	100%	8.6	706,000

Sources
APPSIM - Percival (2007)
CAPP-DYN - Mazzaferro & Morciano (2009), Morciano 2009
CBOLT - PIneles-Mark, Schwabish & Simpson (2009)

CORSIM - Anderson (1997)
Destinie - Blanchet, Buffeteau, Crenner & Le Minez (2009)
DYNACAN - Petrovic (2009)
DYNAMOD2 - Abello, Kelly & King (2002)
DYNASIM - Favreault & Smith (2004)
INAHSIM - Inagaki (2011)
MICROHUS–- Klevmarken & Olovsson (1993)
MIDAS-BE - Dekkers et al (2009), Dekkers (2011)
MOSART - Fredricksen, Knudsen & Stolen (2011)
Moses - Wu, Birkin & Rees (2009)
NEDYMAS - Nelissen (1993)
Orcutt - Orcutt, Greenberger, Korbel & Rivlin (1961)
PENSIM2 - Emmerson, Reed & Shephard (2004)
POLISIM - McKay 2003
PRISME - Albert, Berteau-Rapin & Di Porto (2009)
SAGEMOD - Evandrou, Falkingham, Johnson, Scott & Zaidi (2003)
SESIM - Flood, Jansson, Pettersson, Sundberg & Westerberg (2005), Petersson (2009)
Sfb3 - Galler & Wagner (1983)
SMILE - O'Donoghue, Lennon & Hynes (2009)
SVERIGE - Holm, Holme, Makilla, Maffsson-Kauppi & Mortvik (2004)

More comprehensive lists of models, and reviews of microsimulation methods, are in O'Donoghue (2001), Ahmed and O'Donoghue (2007) and Dekkers et al (2009).

2.4 Cross-sectional, closed, dynamic, discrete-time models

All the models in table 2.1 are "cross-sectional", in that they model all the persons alive at a particular time, and the interactions between these persons. All are "closed", in that the only new persons are births or immigrants, and potential spouses are selected from other living individuals in the data set. In an open model, a synthetic person is created as a partner for each person selected to marry. A similar issue arises in finding dwellings for households. The British model Moses uses a stock of houses that are independent from the households that occupy them (Wu, Birkin & Rees 2009 p4). By contrast, the Australian model APPSIM simulates the value of a dwelling bought by a moving household, rather than selecting from available dwellings (Kelly & Keegan 2008 p21).

All the models in table 2.1 are "dynamic", in that they update each attribute of each person in each time interval. Caldwell (1983 p449) commented

> *"Cost constraints in policy analyses often favor static over dynamic microsimulations. But I believe that dynamic ageing is the more attractive approach. Static aging often rests on hidden behavioral assumptions, and hidden assumptions are inevitably less realistic."*

Static ageing was described by O'Donoghue (2001 p9) as

> *"a procedure that takes a sample whose underlying characteristics X_i are held constant, while the weights given to different parts of the sample are changed through the use of a dynamic reweighting mechanism to produce different weighted distributions corresponding to expected characteristics in the future".*

Noting the inefficiencies and inflexibility of static ageing, he concluded

> *"Static ageing procedures are relatively well suited to short to medium term forecasts, of approximately 3-5 years where it can be expected that large changes have not occurred in the underlying population."*

Given the limitations of static ageing, it is surprising that this technique was adopted for a 20-year projection model of chronic diseases in Australia (Walker, Butler and Colagiuri 2008 p23).

Nearly all the models in table 2.1 use discrete-time modelling, where all the properties of each entity are updated each time-cycle. An alternative approach is the use of continuous time to explicitly model the length of time to events changing the state of individuals. All modelled events are stored in time order, and implemented as their time is reached. A recent example is Charette and Gribble (2009), who modelled the propagation of a hypothetical disease through a social network of hosts. The use of continuous time introduces additional complexities, and run times that may be proportional to N^2 or NlogN, where N is the number of persons modelled.

The Cumpston model uses discrete time modelling, except that the incidence and development of diseases is modelled on a continuous time basis within simulation cycles. This allows for the rapid development of some disease stages, such as terminal cancers.

2.5 Links between national microsimulation models

The first computer-based national household microsimulation model was created by Orcutt, Greenberger, Korbel & Rivlin (1961). In 1969 work began on DYNASIM at the Urban Institute, Washington DC, under the direction of Orcutt. The first version of DYNASIM was completed in 1975 (Orcutt, Caldwell and Wertheimer 1976). A recent version of DYNASIM is described by Favreault & Smith (2004).

CORSIM began its development at Cornell University in 1987. Strategic Forecasting was founded in 1994 by Caldwell, to continue development of CORSIM. In 1995 CORSIM was used as the basis for DYNACAN, with US data, equations and regulations being replaced by their Canadian counterparts. In 1997 CORSIM was used as the starting point for SVERIGE, which itself served as the template for the development of a Chinese model in 2001. In 2001 the US Social Security Administration purchased CORSIM, with Strategic Forecasting providing 3 years of technical assistance during the development of POLISIM (Strategic Forecasting 2002).

Holm, Holme, Makila et al (2002 p3) noted that CORSIM was one of the main points of departure for SVERIGE. The version of CORSIM completed in August 1995 had 700 distinct equations representing 25 equation-based processes. The behavioural content of these equations was “Swedenised”, but CORSIM’s simulation engine was not used.

The US Social Security Administration found that sections of the CORSIM code were hard to understand because they were so dense. Many parts of the code were tuned for efficiency, at the expense of readability. They spent much time rewriting the code in C++, adding documentation. The new code was easier to understand and maintain, but perhaps 50% slower (McKay 2003 p3).

The Life-Cycle Income Analysis Model (LIAM) is a flexible computing framework intended as a basis for dynamic microsimulation models. The principal computer characteristics include modularization, parameterization, generalization and robustness (O’Donoghue, Lennon & Hynes 2009). It has been used in

- Tax analyses with the Irish Dynamic Cohort Microsimulation Model
- Dynamic microsimulation of expenditures in the EUROMOD framework
- Simulation Model of the Irish Local Economy (SMILE)

- EU 6th Research Framework Programme, Adequacy of Old Age Income Maintenance in the EU (AIM).

Simulations for Belgium, Germany and Italy, done as part of AIM, are described by Dekkers, Buslei, Cozzolino et al (2009).

Some of the techniques used in early models were adopted uncritically by other models, and subsequently found to be defective. For example, a "stable marriage algorithm" was used in CORSIM, and subsequently adopted by DYNACAN, POLISIM and SVERIGE. Although Pareto optimum, this algorithm gave unrealistic results (Bouffard, Esther, Jonson, Morrison & Vink 2001 p6).

2.6 High costs of national microsimulation models

Krupp (1983 p37) noted the very high costs of microsimulation models, arising from the need to have a team of

> *"highly-qualified scholars, who in addition to being thoroughly versed in the substantive areas to be studied are also trained in computer programming"*

Harding (2007 p4) noted that the developers of the MINT, CBOLT and POLISIM models of the US had estimated that the total development costs of each model had exceeded US$6m. She considered that budgets of this magnitude had also been involved for the Canadian DYNACAN model, the Norwegian MOSART model and the Swedish SESIM model. The UK SAGE project had five years of grant funding of £0.8m, and NATSEM's 5-year grant for APPSIM was A$1.7m. Harding commented

> *"Today, I would place much greater importance on developing the simplest possible (but functioning) version of a model … within project budget and timeframe … Such an approach militates against the taking of risk, which was a feature of NATSEM's original DYNAMOD model …"*
>
> *"One possible solution to this 'risk' dilemma is for PhD students to undertake the high-risk development of new methods and innovations …"*

Harding, Baldini and O'Donoghue are some of the postgraduate students who have constructed national models as part of their PhDs (O'Donoghue, Lennon & Hynes 2009). PhD students do not have to meet client needs, can limit data analysis, and can omit alignment or validation. But PhD students are subject to stringent external examination, a valuable quality control process.

2.7 Uses for national models

Some examples of the objectives of current models

- The aim of APPSIM is to assist in assessing the future distributional and aggregate social and fiscal implications of Australian government policy (Kelly 2007 p3)

- CAPP-DYN allows analysis of intragenerational and intergenerational effects of reforms in the Italian pension system (Mazzaferro and Morciano 2009 p2)

- DYNAMITE's principal aim is to study the effects of demographic ageing and the evolution of family structures on aggregate saving in Italy (Dekkers et al 2009 p17)

- DYNASIM3 analyzes the long-run distributional consequences of retirement and ageing issues, simulating social security benefits, pensions, home and financial assets, and health status (Favreault & Smith 2004 p1)

- INAHSIM is being used as a comprehensive microsimulation model for policy simulation in Japan, including the effects of pension reform on the income distribution of the elderly (Inagaki 2009 a & b)

- MIDAS is being developed to simulate the adequacy of pensions in Italy, Germany and Belgium (Dekkers et al 2009 p18)

- MOSART is being used to project demographic development, educational performance, labour supply and pension expenditures in Norway for the Ministries of Finance and Labour, and other public institutions, private organisations and media use the results occasionally (Fredriksen, Knudsen & Stolen 2009 p2)

- Moses is intended to provide a generic framework for social scientists and decision makers with a shared interest in modelling and the simulation of social systems within an urban environment (Wu, Birkin & Rees 2009 p3)

- The principal purpose of Pensim2 is to estimate the future distribution of UK pensioner incomes, thus enabling analysis of the distributional effects of proposed changes to pension policy (Emmerson, Reed & Shephard 2004 p1)

- POLISIM has been used for actuarial and statistical estimates for the US Social Security Administration (McKay 2003 p1)

- SESIM was developed in 1997 as a tool to assess the Swedish education financing system, and has since been extended to include the financial sustainability of the Swedish pension system and health issues amongst the elderly (Flood et al 2005 p5).

Excluding Moses, these models are largely being used for long-term national purposes, with little or no need for regional subdivisions. These limited purposes help explain the limited capabilities of most national models to deal with fine geographic scales or short time frames.

O'Donoghue (2001 p17) noted that

> *"Large cross-sectional models … usually simulate a wide variety of economic and demographic processes and can therefore be used for many different applications."*

Progress to date has however been limited. O'Donoghue, Lennon & Hynes (2009 p16) commented

> *"The models that are being used at present are not doing very much more than the DYNASIM model in the late 1970s. A large potential reason is the apparent benefit to cost ratio."*

Most of the objectives of present national microsimulation models appear to be related to the long-term social and fiscal effects of government policy. This narrow range of uses may reflect their general lack of regional detail, long simulation cycles and slow run times. A major objective of this thesis was to find ways to overcome these problems, and create models with a range of government, business and individual uses.

Wolfson (2009 p26) commented

> *"…microsimulation models are still lacking the full range of agents in the economy as envisaged by Orcutt, particularly firms. And aside perhaps from marriage in dynamic models, individuals do not interact with each other; there is nothing like the standard economic notion of markets"*

This failure to achieve Orcutt's objectives may partly reflect the much expanded use of general equilibrium models for macroeconomic projections. Ahmed and O'Donoghue (2007) provide references to many examples of linkages between microsimulation and general equilibrium models.

2.8 Demise of some national microsimulation models

Gupta and Harding (2007 p10) listed the following as dead, or replaced by more complex models

- The German Sfb3 model
- The Australian HARDING model
- The Canadian cohort model DEMOGEN
- The Swedish MICROHUS model.

In addition, the following are not operating, or have never operated routinely

- CORSIM is no longer available, as Strategic Forecasting has ceased operation
- DYNAMOD appears never to have operated routinely, although many of the lessons learned have been valuable in the development of APPSIM
- DYNACAN ceased operation on 1 June 2009, after operating routinely since about 1999
- SAGE was developed primarily as a research tool, and enhancements may form the basis of PhD study (Evandrou et al 2003 p447)
- SVERIGE has never operated routinely, although it continues to be a test platform.

High data maintenance costs, and limited potential uses, may have contributed to their demise.

2.9 First documented uses of key household modelling techniques

Table 2.2 shows the crucial influence of Guy Orcutt in suggesting and then implementing key household microsimulation techniques. Although batch matching is attributed above to Orcutt, Caldwell and Wertheimer (1976), it is likely that this was implemented in earlier models.

Table 2.2 First documented uses of key household modelling techniques

Technique	First documented use
Regional migration	Hagerstrand (1957)
Household microsimulation	Orcutt, Greenberger, Korbel & Rivlin (1961)
Dynamic models	Orcutt, Greenberger, Korbel & Rivlin (1961)
Cross-sectional models	Orcutt, Greenberger, Korbel & Rivlin (1961)
Short simulation cycles	Orcutt, Greenberger, Korbel & Rivlin (1961)
Closed models	Orcutt, Greenberger, Korbel & Rivlin (1961)
Immediate matching	Orcutt, Greenberger, Korbel & Rivlin (1961)
Event alignment	Orcutt, Greenberger, Korbel & Rivlin (1961)
Event listing	Orcutt, Greenberger, Korbel & Rivlin (1961)
Batch matching	Orcutt, Caldwell & Wertheimer (1976) ?

2.10 Immediate versus batch matching

The marriage process of Orcutt et al (1961) was “immediate”, in the sense that a single person was selected to marry, and a spouse found for them as soon as possible. By contrast, Orcutt, Caldwell and Wertheimer (1976 p34-35) selected males and females to be married during a year, then partnered them in a batch process at the end of the year.

Leblanc, Morrison & Redway (2009) tested several existing and proposed batch matching procedures, and concluded that the stochastic batch algorithm documented by Bouffard et al (2001) generated marriages that best fitted the census data. They were however concerned about the validity of their compatibility measure.

The analyses in chapter 6 of this thesis suggest that batch matching will generally give inferior results to immediate matching. When using short time cycles and many regions, some form of immediate matching becomes essential. The Cumpston model uses immediate matching throughout, for partnership formations, entrants to households and matching households to dwellings.

2.11 Event alignment

Orcutt et al (1961 p294-299) described a “tracking mechanism” used to keep births, deaths, marriages and divorces in line with actual past numbers or expected future numbers. Each simulation period, event probabilities were adjusted by factors intended to correct any past excesses or shortfalls of that event. Morrison (2006 p4) notes that such linear adjustments could yield probability values lying outside the range [0,1]. The monthly simulation cycles used by Orcutt et al may have avoided this problem, as the event probabilities would all have been low, and the adjustment factors close to unity.

Morrison notes that improvements in computer technology made it increasing feasible to explicitly recognize alignment pools and adopt two-pass approaches. The most

widely used of these, alignment by sorting, was developed by Johnson (2001). After generating the events for the pool, sufficient of those instances where the random numbers are closest to the a priori probability are reversed to give the alignment total for the pool. In determining "closeness", both the random number and the probability are transformed using ln((1-x)/x).

Chenard (1998) described a complex procedure used by DYNACAN to ensure that the expected numbers of persons migrate from one region to another, with the correct sexes and age groups, and with realistic family structures. This complexity appeared to be needed because potential migrants were presented in non-random order, and because very little deviation from the alignment totals was allowed.

A single-pass alignment procedure was suggested by Cumpston (2009), involving the simulation of the event for randomly selected persons in a pool, ceasing when the desired event total is achieved. This was implemented in the Cumpston model.

2.12 Error detection and model validation

Orcutt et al (1961 p303) included a routine to list household changes, presumably for checking. Their ability to carry out extensive experiments (p393) was

> *"…frustrated by the loss of a substantial amount of computing time through failure on our part to catch a very obscure programming error at an early stage."*

Caldwell (1993 p506) noted that

> *"Because DYNASIM was expensive to construct and operate, the model was little used for policy exploration. The amount of validation done on the original model, relative to the amount required for achieving policy level credibility, was also inadequate."*

Caldwell and Morrison (1998, p200) considered that credibility and usefulness for microsimulation models had been slow in coming, and that this was in part due to the widespread failure of microsimulation modellers to validate extensively and effectively. They gave examples where CORSIM or DYNACAN outputs had been compared with administrative data or actuarial projections.

Morrison (2008) described the spectrum of techniques used to validate the DYNACAN starting database, model equations and parameters, and also the operation of the various models individually and collectively. In spite of this extensive checking, a misinterpretation of mortality coefficients was masked by alignment procedures, and remained undetected for a time (p20). Some of the CORSIM and DYNACAN validation techniques have been applied to the Cumpston model.

2.13 Computer techniques

Wertheimer, Zedlewski, Anderson & Moore (1983 p189) noted that DYNASIM made heavy use of random access disk files, while processing in DYNASIM2 was primarily sequential, with only minimal use of random access files. While this helped reduce runtimes, it appears to have also limited the types of simulation technique used. Several of the simulation, matching and alignment techniques suggested in this thesis require random access to data.

Muller (1983 p496) was concerned that most microsimulation programs used sequential file access methods, rather than data base management systems. This required analysts to take into account the navigational procedures to route through complicated physical structures. Data integrity was also an issue

> *"Ideally, integrity should be controlled by a data management authority and not by the programming analysts who should concentrate on model objects rather than on physical data access".*

One issue in writing the Cumpston model was whether to use object-oriented programming. Yevick (2005 p5) noted that

> "Features such as objects and class introduce complex programming syntax. Accordingly, an object-oriented language requires far more time to learn than a procedural language … Object-oriented development is … often said to be required by programs with over 250,000 lines".

Farooq, Salvini & Miller (2008) describe the object-oriented structure used for the ILUTE model of Greater Toronto. Leaving aside their traffic simulations, most of the 50 ILUTE classes have functional parallels in the Cumpston model, and many of the attributes and operations of each class also have parallels. Object-oriented programming should not affect functionality or execution speed, but may ensure safer development in a multi-programmer environment. It was not used for the Cumpston model.

2.14 Technical details and run times for national models

Table 2.3 gives standardised runtimes for a number of current and old models.

Table 2.3 Run times for national models

Model name	Language	Lines code	Cycle length	Years projected	Runtime in mins	Mins per m years
APPSIM	C#		year	50	1440?	110
CAPP-DYN	STATA	>6,000		45	180	23
CBOLT	Fortran90		year	75	20	0.5
CORSIM	C			30	240	38
Cumpston	C#	13,000	>=day	50	1.42	0.12
DYNACAN	C/C++			130	20	0.48
INAHSIM	Fortran90			100	0.5	0.06
MIDAS-BE	C#		year	10	9.4	4.7
MOSART	C#		year	100	40	0.05
Moses	Fortran			30	150	5.7
Orcutt			month	1	65	6500
Pensim2	Excel/SAS	141,000	year	60	720	180
POLISIM	C/C++	150,000		130	180	3.2
PRISME	SAS	6,200	3 months	41	300	2.0
Sage	C++?		year	40	7	2.9
SESIM	VB6	33,500	year	50	10	1.7
SVERIGE	C++?		year	10	15	0.17

Sources
APPSIM - Kelly (2010)
CAPP-DYN - Morciano (2009)
CBOLT - Sabelhaus (2009)
CORSIM - Anderson (1997)

DYNACAN - Petrovic (2009)
INAHSIM - Inagaki (2011)
MIDAS-BE - Dekkers (2011)
MOSART - Fredricksen, Knudsen & Stolen (2011)
Moses - Birkin (2009)
Orcutt - Orcutt, Greenberger, Korbel & Rivlin (1961)
Pensim2 - Emmerson, Reed & Shephard (2004)
POLISIM - McKay 2003
PRISME - Albert, Berteau-Rapin & Di Porto (2009)
SAGEMOD - Evandrou, Falkingham, Johnson, Scott & Zaidi (2003)
SESIM - Flood, Jansson, Pettersson, Sundberg & Westerberg (2005), Pettersson (2011)
SVERIGE - Holm, Holme, Makilla, Maffsson-Kauppi & Mortvik (2004)

Standardised runtimes have been calculated in table 2.3 as minutes per million person years. For example, INAHSIM takes about half a minute to project forward 100 years, starting with 128,000 persons and ending with about 47,000. This gives

0.5 / (100 * (0.128+0.047)/2) ie about 0.06 minutes per million person years.

Where the end population was not available, future growth was assumed to be at the annual growth rate from 2008 to 2050 projected by the Population Reference Bureau (www.prb.org). The run time for the Cumpston model is for a yearly simulation cycle, including disease simulation.

2.15 Reasons for differences between run times

Table 2.3 shows strong variations in standardised run times, with a range of about 3,600 to 1 between the slowest and fastest recent model. Possible reasons for these variations include

- different functionality, such as internal migration
- inclusion or exclusion of non-demographic events (the run time quoted for DYNACAN excludes non-demographic events such as employment)
- inclusion or exclusion of data input and results output (the time for the Cumpston model includes data input but excludes detailed data outputs)
- different computers (even faster run times would be likely for SVERIGE using modern computers)
- different simulation periods (PRISME has a quarterly simulation cycle)
- different simulation languages (simulations done in multi-purpose statistical packages such as SAS and STATA are likely to be slower)
- uncertainty about the run times of models under development, such as APPSIM
- alignment to external totals during projections (times for INHASIM and the Cumpston model include no alignment, but the times for other models such as DYNACAN may include extensive alignment)
- different simulation techniques (the Cumpston model uses sampling with loaded probabilities, which should be quicker than the all-cycle simulation used by other national models)
- different matching techniques (the Cumpston model uses “best of n” matching, which should be quicker than the matching techniques used by most other national models).

Validation of APPSIM was due to be completed in June 2011, and improvements to its run times were deferred until completion of validation (Kelly 2010).

Very long run times may reflect inadequate memory or slow data access, as well as excessive alignment.

Klosgen (1983 p490) noted the potentially large speed improvements available from multithreading, but this technique does not appear to have been used to any great extent. SVERIGE was unable to get multithreading to work reliably (Holm, Holme, Makila et al (2002 p3). Parts of POLISIM are multithreaded, but this required careful code construction to avoid crashes (McKay 2003 p3). MOSART has recently been multithreaded, with substantial runtime reductions (Fredriksen, Knudsen & Stolen 2011 p5-7).

2.16 Advantages of short run times

Modern computers are so fast that there is normally little need to worry about reducing run times. But household microsimulations can involve very large amounts of data, complex calculations and surprisingly long run times.

Faster run times may allow

- Finer geographical subdivisions
- Higher sampling densities
- Modelling of events at time scales as short as days
- Multiple runs to quantify uncertainties in projected outcomes (see for example Simpson & Topoleski 2005)
- Searches for assumptions better matching available data
- Faster debugging and validation
- Faster responses to new requests
- Iterative linking to macroeconomic models
- A wider range of government and commercial applications.
- Higher staff productivity.

2.17 Advantages of short simulation cycles

Orcutt et al (1961 p26-27) said in relation to their model

> *"An essential characteristic of this type of model is that it proceeds in short discrete steps or periods. It is a recursive model: that is, the outputs of a unit in any one period depend on prior inputs to the unit, so that there is no simultaneous interaction between units, and hence no simultaneous equations involving more than one unit of time to be solved."*

This is in contrast to the many simultaneous equations requiring solution in the general equilibrium models then emerging, for example in Johansen (1960). Unlike the explicit time steps of microsimulation, the general equilibrium models did not operate on any defined time scale.

Orcutt et al continued

> *"If a recursive model is to be a good approximation to reality, the time period chosen must be short… We have used the month as our basic period. An even shorter period, such as the week, might sometimes be preferable…"*

Galler (1997) looked at discrete-time and continuous-time approaches to dynamic microsimulation, and concluded that a discrete-time framework with comparatively short time cycles appeared to be best suited for causal models.

In spite of the theoretical advantages of short cycles, most of the current national household microsimulation models operate on yearly cycles. One exception is PRISME, which uses 3-month cycles to allow for the wide use of 3-month periods in French legislation (Albert 2009).

With one-year cycles, the order in which events are simulated can cause appreciable differences in the results. For example, immigrants add about 1.7% to Australia's population each year, and may provide about 3.2% of national births in the year after they arrive. Modelling births before immigrants in a yearly cycle may thus lose about 1.6% of births (assuming that immigrants arrive evenly throughout the year).

Analysis shows that the size of such errors reduces linearly with the length of the simulation cycle. For example, a quarterly simulation cycle would reduce the error from modelling births before immigration to about 0.4% of births. Shorter cycles can also reduce more subtle errors - for example, modelling immigrants with a yearly cycle can exhaust the supply of vacant dwellings, or have an undue impact on simulated dwelling prices.

Some events require short time cycles to be realistically simulated. For example, persons trying to sell their dwellings may decide whether to accept any bids, or change their asking price, on a weekly basis. Similarly, persons trying to buy dwellings may bid for dwellings, or review their requirements, on a weekly basis. Persons travelling within a city may decide their travel plans at least daily. The Cumpston model can use simulation cycles as short as a day.

2.18 Advantages of fine geographic subdivisions

National or sub-national models with more geographical subdivision or shorter simulation cycles could have a wide range of uses, provided their run times can be kept reasonable. For example

- Models with fine geographic subdivisions could be used to select viable locations for long-term facilities (such as schools and age-care facilities)
- Models with fine geographic subdivisions and short time cycles could be used to simulate housing markets
- Models with very fine geographic subdivisions and very short time cycles could be used to simulate metropolitan traffic flows, as in the ILUTE project for Toronto (Miller 2009a).

Caldwell (1983 p61) commented

> *"Policy analysis payoffs become more substantial when spatial disaggregation reaches the state level in American models. But the gains accumulate exponentially when the level of spatial disaggregation becomes even finer than the state level."*

Discussing the need to model interactions between individuals and firms, he said

> *"Both corporate and individual actors necessarily operate within settings - constraints and opportunities - that condition their behavior … Since many important constraints are local … models should aim to represent contexts in a spatially disaggregated manner."*

Caldwell commented (p73)

“The major difficulty with precise location identification is the need for far larger samples in order to reduce sample variance for area-specific outcomes. One might imagine working with one percent, five percent or even ten percent samples of the population.”

The SVERIGE model of Sweden (Holm et al 2004) modelled all 8.6m Swedes, spread over 108 labour market regions, and further located into 706,000 squares of 100 by 100 metres. The Moses model (Wu, Birkin & Rees 2009, Birkin 2009) models 750,000 persons in 33 local areas within Leeds, and is developing this fine-area capability for the 61m persons in the UK. The SVERIGE model of Sweden models 9 regions and 290 municipalities (Pettersson 2011). The Cumpston model locates 175,000 persons over 57 statistical divisions.

2.19 Data synthesis for small areas

Voas & Williamson (2000) recommended the reweighting of survey data using combinatorial optimization for synthesis of small area data. Birkin and Clarke (2009) compared this approach with the earlier use of chain imputation:

"Early applications of spatial microsimulation constructed data sets through conditional probability models where households were located first within a city or region and then household attributes were built up for that household based on Monte Carlo simulation. This involved the use of random number tables to determine whether the household should be given certain attributes or not based on the household location ... This methodology had the advantage that the geographical patterns produced were very robust, in the sense that the household began with a geographical location and added variables in direct relation to the information contained in the local census tables."

"However, it was not easy to reproduce (or certainly test) the interdependencies between variables within households, especially when non-census variables were introduced. In the mid 1990s therefore the favoured methodology shifted towards the reweighting of survey data using combinatorial optimisation procedures ... it is becoming increasingly apparent that the geographical patterns ... can look 'too flat': that is, households that are least likely to be in the surveys (perhaps households of extreme affluence and poverty) are very difficult to reproduce across a city or region..".

Cumpston (2007) suggested a sequential allocation method for small area data synthesis. This is similar to chain imputation, except that the numbers of each value of a characteristic for a small area are allocated rather than randomly generated. The numbers of persons of each sex and age in each small area are input data. The numbers of persons with each value of a characteristic for each sex and age group are known from small-area cross-tabulations. The probabilities of a particular value being associated with the other characteristics known at that stage of the synthesis are derived by logistic regression of large area data. These probabilities are calculated for each eligible person in the small area, and the person receiving the particular characteristic selected by probability-weighted sampling. Persons in the small area are grouped into households, similarly using logistic probabilities of association. A version of this method was used to synthesize households from Australian census data for statistical local areas (Sarjeant 1998).

This sequential allocation procedure has some advantages. All marginal totals for each characteristic are correctly replicated by sex and age group for each small area. As no form of optimisation is involved, no search failures can occur. The process is robust, and should be able to cope reasonably with data confidentialisation and non-

responses. Checks on the validity of the allocation process can be obtained by comparing cross-tabulations of pairs of characteristics. This proposal is not further discussed in this thesis, as it is a major research topic in its own right.

3. SIMULATION BY SAMPLING WITH LOADED PROBABILITIES

All-case simulation, where demographic outcomes are simulated for each person each year in file order, has been used in almost every national household microsimulation model. Cumpston (2010a) suggested the use of simulation by sampling with loaded probabilities, where carefully chosen numbers of cases are drawn from pools, and Monte Carlo simulation applied with appropriately loaded probabilities. This suggestion is intended to provide faster simulations, particularly when using simulation cycle times shorter than a year.

Formulas are derived for optimal draw numbers and loaded probabilities, and found to be of forms which can be implemented in household microsimulations. Different formulas are needed for events which do not cause losses from a pool, such as movements within a geographic pool, and for events such as deaths which do cause losses.

The formulas for events with pool losses are tested by repeated trials with pools containing cases with only two risk levels. More realistic testing, using the Cumpston model, is done to look at

- One-year runtimes using all-case and loaded sampling, with simulation cycles ranging from yearly to daily
- 50-year projections with all-case sampling with yearly cycles, and with loaded sampling with yearly and weekly cycles
- Simulated one-year event numbers with all-case sampling with yearly cycles, and with loaded sampling with yearly and weekly cycles
- Event numbers with loaded sampling when simulating in normal and reverse order, with yearly and weekly cycles
- Standard deviations of event numbers, with all-case and loaded sampling.

Conclusions from this testing, and comments on the practical applicability of sampling with loaded probabilities, are at the end of the chapter. In summary, sampling with loaded probabilities is theoretically valid, works in practice, and can reduce runtimes.

3.1 All-case simulations

The first household microsimulation model developed in response to Orcutt's 1957 proposals simulated demographic outcomes for each person in turn, working in the order in which the persons were stored (Orcutt, Greenberger, Korbel & Rivlin 1961). The limited memory capacity of the available computers would not have been able to hold the 10,000 persons simulated, and it is unlikely that random access to external storage media was feasible. Simulating outcomes for each person in file order would thus have been natural. This process has apparently been followed in every subsequent national model.

3.2 Simulation by sampling with loaded probabilities

Testing every person for the occurrence of a demographic event seems inefficient, particularly for low-probability events. As an alternative, can we divide persons into pools, and only test a small proportion of persons in those pools with low probabilities? The assumed probabilities of the event occurring would have to be loaded, to offset the partial sampling. Such a process might be particularly efficient in short simulation cycles, where the probabilities of events occurring in the cycle are very low.

3.3 A trivial case: uniform event probabilities

Suppose we have 100 persons in a pool, each with 20% chance of death. To simulate the expected 20 deaths, we can randomly select any 20 persons from the pool, and declare them to have died. This simple procedure will rarely be applicable, as the expected number of events from a pool is unlikely to be an integer. For example, if there are 101 persons in the pool, each with a 20% chance of dying, then 20.2 deaths are expected. To obtain simulations with the correct expected number of deaths, we can randomly draw 21 persons from the pool, select a random number between 0 and 1 for each of them, and declare them to be dead if the random number is less than 20.2/21 (ie 0.9619). The number of draws needed in this case is reduced by 79%, increasing computational efficiency.

More formally, if there are n persons in a pool, each with probability p_i of an event occurring, then the number of draws d needed from the pool is

$$d = p_i\, n \qquad \text{to the higher integer} \qquad (3.1)$$

Let q_i be the loaded probability of the event occurring for each of the drawn persons. Then

$$q_i = (p_i\, n) / d \qquad (3.2)$$

Substituting $p_i = 0.2$, $n = 101$ and $d = 21$ gives $q_i = 0.9619$, as above.

As each person has the same probability of the event occurring, this procedure works for events such as death where a person is lost from the pool, or events such as moves where the person remains within the pool.

3.4 Non-uniform event probabilities, with no losses from pool

The previous section assumed uniform probabilities. If persons in the pool have varying probabilities of the event occurring, a more complex loaded sampling procedure is needed. Suppose that the highest probability is p_{max}. From equation 3.2, restricting the maximum loaded probability to not exceed 1 requires that

$$(p_{max}\, n) / d \;<= 1 \qquad (3.3)$$

so that the number of draws needed from the pool is

$$d = p_{max}\, n \qquad \text{to the higher integer} \qquad (3.4)$$

As before, the loaded probability of the event occurring for person i is given by equation 3.2. This analysis is appropriate for events not causing losses from a pool - for example, movements within a geographic pool.

3.5 Non-uniform event probabilities, with losses from pool

The previous section assumed no losses form the pool, but some events, such as deaths and emigration, will cause losses. As sampling proceeds from the pool, the cases most at risk are likely to be selected, and the risk profile within the pool will

change. This requires some modifications to the formulas for loaded probabilities and draw numbers.

Let n_t be the number of persons in the pool immediately before the t^{th} draw. Also let q_{it} be the loaded probability for person i during the t^{th} draw. The probability of any person being picked in the t^{th} draw is $1 / n_t$, so the probability of person i being simulated to have the event in draw t is q_{it} / n_t. The probability of person i not being simulated to have the event in any of d draws is

$$(1 - q_{i1} / n_1) (1 - q_{i2}/ n_2) \ \ldots \ (1 - q_{id} / n_d) \qquad (3.5)$$

We do not know n_2, n_3 … n_d in advance, because they will depend on whether the preceding draws have removed any persons from the pool. But we can make the probability of person i being selected to have the event equal in each draw, by calculating loaded probabilities for each draw after the first with the formula

$$q_{it} = q_{i1} \ (n_t / n_1) \qquad (3.6)$$

Substituting equation (3.6) in (3.5), and setting the result equal to the desired probability of person i not having the event, gives

$$(1 - q_{i1} / n)^d = 1- p_i \qquad (3.7)$$

Taking the $1/d^{th}$ power of both sides and rearranging gives

$$q_{i1} = n \ [1 - (1- p_i)^{1/d}] \qquad (3.8)$$

Now q_{i1} is the loaded probability of person i being selected for the event in draw 1, and we do not want this to exceed unity. Thus we require, for all the persons in the pool, that

$$n \ [1 - (1- p_i)1/d] <= 1 \qquad (3.9)$$

ie $$(1- p_i)1/d \ >= 1 - 1/n \qquad (3.10)$$

Taking logs of both sides and rearranging gives the requirement that

$$d >= \ln(1 - p_i) / \ln(1 - 1/n) \quad \text{for all i} \qquad (3.11)$$

This can be achieved by taking d as

$$d = \ln(1 - p_{max}) / \ln(1 - 1/n) \quad \text{to the higher integer} \qquad (3.12)$$

where p_{max} is the highest probability of the event occurring for any member of the pool. Note that if p_{max} is small and n large, this formula for d approximates to that in equation 3.4. Once d has been chosen using equation 3.12, it remains unchanged till all the d draws are completed. For the person selected for testing in the t^{th} draw, the loaded probability q_{it} can be calculated by combining equations 3.6 and 3.8 to get

$$q_{it} = n_t \ [1 - (1- p_i)^{1/d}] \qquad (3.13)$$

As the number n_t in the pool cannot increase above its initial value, the loaded probabilities cannot exceed their values for the first draw, and thus cannot exceed 1.

3.6 Numbers of draws when using sampling with loaded probabilities

Table 3.1 shows the numbers of draws needed for pools of varying size, calculated using equation 3.12. The minimum number of draws from any pool is 1, so that sampling from very small pools becomes inefficient. For maximum event probabilities of about 0.63 or higher, the number of draws can exceed the number of persons in the pool. While this may appear inefficient, in practice this is likely to occur for relatively few pools.

Table 3.1 Number of draws to give maximum loaded probability of 1

Number	Maximum unloaded probability						
in pool	0.01	0.02	0.05	0.1	0.2	0.5	0.9
10	1	1	1	2	3	7	22
100	2	3	6	11	23	69	230
1000	11	21	52	106	224	693	2302
10000	101	203	513	1054	2232	6932	23025
100000	1006	2021	5130	10536	22315	69315	230258

3.7 Tests on accuracy

There was some concern that the formulas in section 3.5 might not give completely correct results in practice. To test this, 10,000 repeated trials were made with risk pools containing varying numbers of low risk cases, and 500 high risk cases. In the first set of trials, 2,500 low-risk cases were assumed, each with probability of 0.1, so that 250 events were expected from the low-risk cases. In the second to fifth set of trials, the assumed probability for the low-risk cases was progressively increased, and their number reduced, so as to maintain the expected number of low-risk events at about 250. In each set of trials, the 500 high-risk cases were assumed to have probability 0.5, so that 250 events were expected from them. In each set of trials, about 500 events were thus expected.

Table 3.2 shows that the mean numbers of simulated events were all very close to the expected 500. The worst deviation from the expected 500 events was 0.12, for the second set of trials. In no case was the deviation from the expected 500 events greater than the standard deviations in the last column of the table. These standard deviations are the standard deviations in the observed event numbers from 10,000 trials, divided by 100 (the square root of the number of trials). These results suggest that sampling by loaded probabilities, for the case with non-uniform event probabilities and losses from the pool, can work correctly.

Table 3.2 Simulated event numbers with pools chosen to produce 500 events

Number low risk persons	Low risk rate	Number high risk persons	High risk rate	Expected number of events	Mean number from 10,000 trials	Standard deviation from 10,000 trials
2500	0.1	500	0.5	500.00	499.92	0.16
1250	0.2	500	0.5	500.00	500.12	0.14
833	0.3	500	0.5	499.90	499.88	0.12
625	0.4	500	0.5	500.00	500.03	0.10
500	0.5	500	0.5	500.00	500.00	0.09

3.8 Example of the use of pools in simulating deaths

In practice, there can be considerable heterogeneity in risk rates, so that segregation into a small number of risk groups can greatly reduce the numbers of draws needed to simulate the expected numbers of events. This is illustrated in table 3.3, using Australian data on ages and mortality rates.

Table 3.3 Draws needed to simulate the deaths in a year from 175,000 Australians

Age Group	Persons in group	Average mortality rate	Expected deaths	Maximum mortality rate	Draws
0-	36441	0.00040	14.5	0.00453	166
15-	23593	0.00060	14.1	0.00113	27
25-	25575	0.00087	22.1	0.00138	36
35-	26810	0.00123	33.0	0.00192	52
45-	24156	0.00237	57.2	0.00415	101
55-	16348	0.00642	104.9	0.01315	217
65-	11852	0.01826	216.4	0.03697	447
75-	7706	0.05051	389.3	0.10627	866
85-	2563	0.18238	467.4	0.48496	1701
Total	175044	0.00754	1319.0		3613

The numbers of expected deaths in table 3.3 were estimated by applying 2001-02 mortality rates to a sample of 175,044 persons derived from the 1% 2001 census sample file. The maximum mortality rates are the highest applying to any age in each of the groups. For example, the highest mortality rate for the 0-14 group is for males aged 0, while the highest rate for persons 85+ is for females aged 110. The numbers of draws needed with loaded sampling were calculated from equation 3.12. For example, the number of draws needed from the pool for persons 55-64 was calculated as

$$\ln(1 - .01315) / \ln(1 - 1/16348) \qquad \text{ie} \quad 217 \text{ to the higher integer}$$

These estimates suggest that good simulations of the deaths in a year can be made by 3,613 draws with loaded probabilities, rather than 175,044 draws with true mortality rates. Similar calculations show that if we treat each age separately for males and females, we can get good simulations with only 1,505 draws.

3.9 Pools used to test sampling with loaded probabilities

In order to explore the achievable efficiency gains offered by sampling with loaded probabilities, the technique was applied to the Cumpston model. The pools in this model are each combination of 8 areas, 8 person types and 9 age groups, making 576 pools in all.

These pools were available for alignment purposes, and not specifically chosen for sampling with loaded probabilities. With hindsight, these pools were too fine for sampling with loaded probabilities, and resulted in simulated numbers of exits and moves being a little different from expected. These errors, and ways to minimize them, are discussed in section 3.13.

3.10 Run times to simulate 175,000 Australian for a year

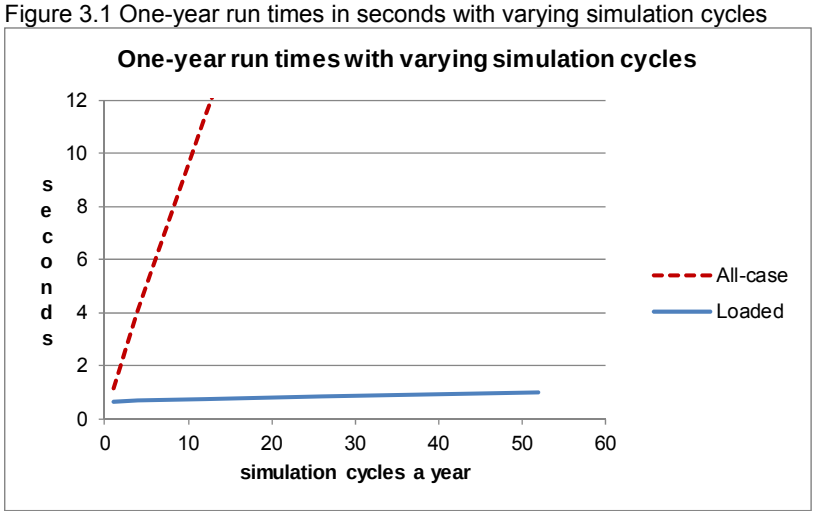
The run times in table 3.4 exclude the input of data and assumptions before a run, and the time needed to output results. There is some random variation in run times, so the run times with yearly cycles are the average of 50 runs, and all the averages for loaded probabilities are the averages of at least 3 runs.

Table 3.4 Mean run times in seconds to simulate 175,000 Australians for a yea

Cycles per year	All-case trials seconds	All-case fitted seconds	Loaded trials seconds	Loaded fitted seconds	Loaded trials as % of all-case
1	1.13	1.58	0.64	0.66	57%
4	4.10	4.14	0.67	0.68	16%
12	11.37	10.98	0.75	0.73	7%
26	23.29	22.94	0.83	0.82	4%
52	44.90	45.15	0.98	0.97	2%
365			2.88	2.88	

Table 3.4 and figure 3.1 show that the times required with all-case simulations increase broadly in line with the number of cycles a year. Fitting a linear trend line by least squares minimization suggests that sampling each case takes about 0.85 seconds a cycle, and adjusting data in response to simulated events takes about 0.72 seconds a year, with a poor fit for yearly cycles. The results with loaded probabilities suggest that sampling takes about 0.006 seconds a cycle, and adjusting data in response to simulated events takes about 0.66 seconds. The fit with loaded probabilities is good up to 365 cycles a year. The numbers of simulated events, and the time taken to process them, should be very similar with both methods. As expected, the time savings are in sampling.

Figure 3.1 One-year run times in seconds with varying simulation cycles



3.11 Run times for projections for up to 50 years

The yearly values in the first line of table 3.5 are those in the first line of table 3.4. They show sampling with loading probability has a one-year run-time about 57% of all-case sampling, which is out of line with the 53% or 52% in the multi-year lines of table 6. Comparing the calculation steps involved in all-case and loaded sampling suggests that the savings from loaded sampling should be a constant proportion, regardless of the number of projection years. The one-year result is an anomaly.

Table 3.5 Mean run times in seconds

Years	All-case Yearly	Loaded Yearly	Loaded Weekly	Loaded as % of all-case Yearly
1	1.13	0.64	0.98	57%
10	11.59	6.18	9.81	53%
20	24.28	12.82	20.28	53%
30	37.35	19.72	31.20	53%
40	51.93	27.25	42.51	52%
50	66.32	34.16	53.87	52%

Figure 3.2 Run times for one-year projections

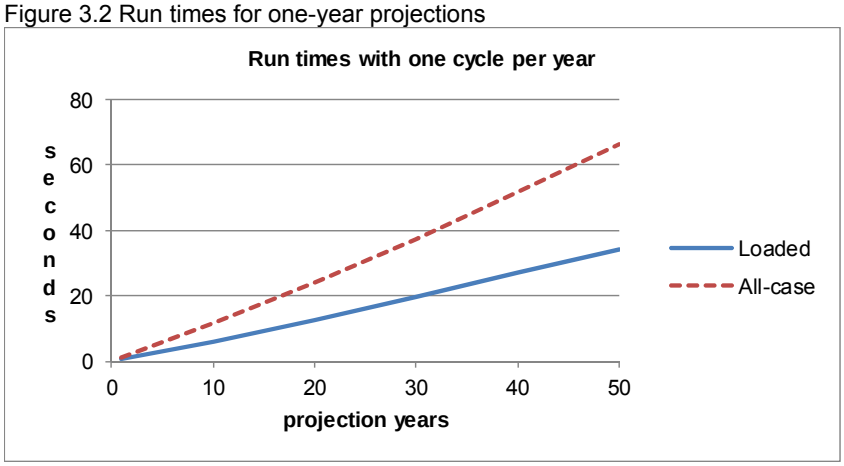


Figure 3.2 shows that both all-case and loaded sampling methods give curves bending slightly upwards. This is because of the population growth of about 30% projected over the 50 years.

3.12 Checks on projected event numbers

Sections 3.13 and 3.14 compare the expected numbers of each type of movement, with the average numbers observed with different cycle time, sampling methods and event orders. The types of event examined are births, deaths, emigrants, exits and moves. Immigrants were not compared, because the microsimulation model pregenerates immigrant families to match exogenous immigrant assumptions. An “exit” is a departure from a household of a person, possibly followed by one or more of the other household members. By contrast, a “move” is where the whole household moves to another dwelling. For various reasons, the observed numbers differ from expected by more than can be explained by random variation. Nevertheless, the comparisons show that broadly reasonable results are being obtained

3.13 One-year projection results with each event simulated separately

Table 3.6 shows the average numbers of projected events from 50 one-year runs, using either all-case simulation or sampling with loaded probabilities. To avoid event interactions confusing the results, only one type of event was simulated in each run. Also shown are the expected numbers of events, obtained by summing the probabilities of each person having that event.

Table 3.6 Event totals from one-year projections, each event simulated separately

Event	Expected	Observed all-case yearly	Observed Loaded Yearly	Observed loaded weekly
Births	2170	2179	2183	2208
Deaths	1372	1379	1330	1370
Emigrants	2284	2280	2251	2275
Exits	6254	6441	6203	6592
Moves	12468	12411	12330	12364

Some of the possible reasons for the differences in table 3.6 are:

- Random variations in the numbers of simulated events. For example, if the expected 1,372 deaths were Poisson distributed, they would have a standard deviation of 37, and averaging over 50 runs would reduce this to about 5. The observed death numbers from all-case yearly and loaded weekly are thus within 1.4 standard deviations, while those from loaded yearly runs are well outside.

- Derivation of the expected numbers in a manner inconsistent with the simulations. Event probabilities, particularly for births, exits and moves, depend on a number of person and household characteristics. While the expected probabilities for each person were based as closely as possible on the data at the start of the first projection year, some approximations may have occurred. One error in expected numbers was found during checking.

- Movements between sampling pools resulting from births, deaths or emigrants. For example, if a partnered male is simulated to die, then his partner will be removed from a partner pool, and put into a pool of lone persons or lone parents. As deaths from these pools are simulated after deaths from partner pools, then the former partner may be double exposed to the risk of death. As birth, death and emigration are low-probability events, and some offsetting may occur, the resulting errors are likely to be small. A solution is to use pools based only on age.

- Changes in type or area for exits. In the simulations with loaded probabilities, it was assumed that exits do not cause any losses from the pool. Trials show that about 86% of exits involve a type change, and about 10% involve an area change. With the area/type/age pool structure being used, changes of either type or area result in removal from the pool being simulated. This is not a problem with all-case simulations, but can cause insufficient persons to be sampled when using loaded probabilities. This seems likely to be a major reason why exits with a yearly cycle and loaded probabilities were 3.7% below the all-case number. A solution is to use pools based only on age, so that no pool losses occur.

- Changes in area for moves. About 14% of moves involve an area change, and moves with a yearly cycle and loaded probabilities were 0.7% below the all-case number. Again, a solution is to use pools based only on age.

- Changes in person types during the year. If type change assumptions are inappropriate, then there can be rapid changes in the numbers of particular types of persons in the first projection year. This may be a major reason for the 6% increase in exit estimates with loaded probabilities when changing from annual to weekly cycles.

- Programming errors. The Cumpston model was primarily constructed as a test-bed for different simulation techniques. It has been substantially debugged and validated, but there is potential for errors of the order of 10% in exit and move numbers to be still undetected, and for smaller errors in births, deaths and emigrants.

3.14 One-year projection results with all events simulated together

Table 3.7 shows the same expected numbers as table 3.6, but with observed numbers obtained by simulating all events together. Within each simulation cycle, events have been simulated in the order births, deaths, emigrants, immigrants, exits and moves. Apart from moves, simulated event numbers are similar for both tables. Move numbers increased by about 8% for all-case simulation, 7% for loaded probabilities with a yearly cycle, and 4% for loaded probabilities with a weekly cycle. Move probabilities depend strongly on person types, so that poorly chosen exit assumptions, or abnormal initial type distributions, can cause significant changes to moves in the first year.

Table 3.7 Event totals from one-year projections, all events simulated together

Event	Expected	Observed all-case yearly	Observed loaded yearly	Observed loaded weekly
Births	2170	2175	2182	2232
Deaths	1372	1375	1335	1375
Emigrants	2284	2282	2263	2304
Exits	6254	6333	6075	6611
Moves	12468	13418	13210	12826

3.15 Comparisons between normal and reverse order simulations

Galler (1997) recommended short simulation cycles. To test their usefulness, simulations were made in normal and reverse order with yearly cycles, and then in normal and reverse order with weekly cycles.

Table 3.8 shows that the changes from simulating events in reverse order are much smaller with weekly simulation cycles, particularly for exits and moves. These results strongly confirm the desirability of short simulation cycles.

Table 3.8 Loaded one-year projections, with events in normal & reverse order

Event	Yearly			Weekly		
	Normal Order	reverse order	change in year	normal order	reverse order	change in year
Births	2182	2243	61	2232	2224	-8
Deaths	1335	1331	-4	1375	1378	3
Emigrants	2263	2308	45	2304	2311	7
Exits	6075	6540	465	6611	6615	4
Moves	13210	12323	-887	12826	12819	-7

3.16 Comparisons between all-case and loaded 50-year projections

Table 3.9 compares the averages of 10 50-year runs using all-case simulation with the averages of 10 runs using loaded probabilities. The expected numbers of births, deaths and emigrants are reasonably similar for both simulation methods. The reasons for the shortfalls in exits and moves are not clear. Exits and moves do not affect person numbers, and the projected numbers of persons are very close.

Table 3.9 50-year projections with yearly cycles

Event/ Persons at end	All-case Yearly Mean	All-case yearly SD	Loaded yearly mean	Loaded Yearly SD	Loaded as % of all-case
Births	102831	541	101611	516	98.8%
Deaths	92205	127	91924	130	99.7%
Emigrants	123775	186	123024	232	99.4%
Exits	355304	891	347052	746	97.7%
Moves	797511	1716	778254	937	97.6%
Persons at end	226755	488	226566	504	99.9%

3.17 Checks on event number standard deviations

Table 3.10 compares approximate expected standard deviations for each event type with those observed from the averages of 50 runs, simulating all events together. For each sampling pool, the average loaded probability was calculated as

expected number of events / number of draws from pool

where the number of draws was calculated using equation 3.4 or 3.12.

Table 3.10 Standard deviations for one-year projections

Event	All-case yearly	All-case yearly	Loaded yearly	Loaded yearly	Loaded weekly	Loaded weekly
	Expected	Observed	Expected	Observed	Expected	Observed
Births	43.8	38.6	43.6	38.7	43.5	52.8
Deaths	35.4	32.7	28.0	29.9	31.7	32.2
Emigrants	48.4	38.6	25.6	35.6	37.5	42.8
Exits	73.7	67.6	33.2	39.8	54.7	59.6
Moves	98.3	99.7	10.3	18.9	50.0	55.1

The expected variance of the number of events from that pool was approximately estimated assuming the variance formula for a binomial distribution

number of draws * average loaded probability * (1 - average loaded probability)

Variances were summed across all pools, and the square root taken to give an approximate estimate of the standard deviation. Given the approximate nature of the expected standard deviations, and the broad confidence limits generally associated with standard deviation observations, the expected and observed standard deviations are reasonably comparable.

As shown in table 3.2, sampling with loaded probabilities can give low standard deviations, particularly for risk combinations with a narrow range of risks. This is because the loaded probabilities can all be close to 1, and there will be little variability in the simulated numbers of events. For moves, most of the persons in each sampling pool had similar probabilities, so that loaded probabilities with a one-year simulation cycle gave standard deviations that were much lower than with all-case sampling.

As the number of simulation cycles in a year increases, more pools have very low numbers of expected events, with loaded probabilities well below 1, and higher variability in simulated numbers of events. Comparing the values for loaded simulations with yearly and weekly cycles, all the observed standard deviations increased, and particularly so for moves.

3.18 Coefficients of variation for one-year projections

The coefficients of variation in table 3.11 were obtained by dividing the observed standard deviations in table 3.10 by the observed numbers in table 3.7. Regardless of simulation method or cycle length, the coefficients of variation for the Cumpston model are small, and may not be important in most practical applications.

Table 3.11 Observed coefficients of variation for one-year projections

Event	All-case yearly	Loaded yearly	Loaded weekly
Births	0.018	0.018	0.024
Deaths	0.024	0.022	0.023
Emigrants	0.017	0.016	0.019
Exits	0.011	0.007	0.009
Moves	0.007	0.001	0.004

3.19 Computational aspects

Computational details for loaded sampling will depend strongly on the database structure and programming language. Test results here are from the Cumpston model, which uses a list structure rather than a relational database. A separate list is maintained of member addresses for each of 576 pools (each combination of 8 areas, 8 person types and 9 age groups). These pools are maintained for alignment purposes, but they also proved useful for loaded sampling. All data for a person are stored as a single line in an array, and their address in the array is recorded in the relevant pool list. The address of each new person in a pool is added to the end of the list. The address of any exiting person is replaced in the list by the address currently at the end of the list, with the length of the list being reduced by one. A random number is drawn, and multiplied by the length of the list, to randomly select a person from a pool.

Sampling pools need to be chosen so as to avoid undesired transfers into other pools. For example, pools used for simulating deaths should not use pools stratified by type of relationship within the household.

3.20 Likely applications

In practice, where is sampling with loaded probabilities likely to be useful? Models with fine geographic subdivisions will generally require at least 1,000 persons to realistically represent each subdivision, and may thus become very large. Models with many different types of disease will also need to be large, so as to adequately represent uncommon diseases. Models including processes with short time spans are also likely to get useful time savings from sampling with loaded probabilities. For example, a microsimulation model of dwelling sales and rentals is likely to need both a detailed geographic structure, and a simulation cycle of a week or less.

Existing models of moderate size but unusually slow run times are unlikely to benefit from sampling with loaded probabilities. Their slowness may reflect programming and data storage issues, or excessive use of alignment, and more efficient sampling will have little effect. Adding sampling with loaded probabilities to any existing model may require changes to data indexing procedures, and will require additional model validation.

Fredriksen, Knudsen & Stolen (2011) describe the use of multithreading to greatly reduce runtimes in the MOSART model of the whole Norwegian population. They comment that simulation steps involved in household formation are cumbersome to multithread, due to often subtle interactions between individuals, with little or no effect on runtime. Sampling with loaded probabilities seems particularly relevant to births, deaths, emigration and household changes, all of which are likely to be hard to multithread.

3.21 Conclusions

Repeated trials with pools with only two risk levels gave results very close to those expected from the formulas for pools with losses.

Tests with the Cumpston model show that sampling with loaded probabilities can reduce demographic run times with yearly cycles by about 43%, and run times with weekly cycles by about 98%. A 50 year demographic projection took 34 seconds with a yearly cycle, and 54 seconds with a weekly cycle.

Numbers of simulated events were reasonably close to those expected from underlying probabilities, for all-case sampling with yearly cycles, and for sampling with loaded probabilities with yearly and weekly cycles. A number of reasons for differences between expected and simulated numbers are suggested. The sampling pools used for the tests appear to be too fine, and pools based only on age may give better results.

Yearly simulations in normal and reverse order gave noticeably different event numbers, while weekly simulations appeared not to be affected by event order. This confirms Galler's 1997 recommendation of short cycles.

Sampling with loaded probabilities generally gives lower event number standard deviations than all-case simulation.

Sampling with loaded probabilities seems appropriate for new ambitious models, with many persons or short simulation cycles. It is particularly relevant for demographic processes, where multithreading is generally not applicable. Sampling with loaded probabilities is theoretically valid, works in practice, and can reduce runtimes.

4. EVENT AND STATE ALIGNMENT

Microsimulation projections are often required to align with external totals. Event alignment was first used by Orcutt, Greenberger, Korbel & Rivlin (1961). Subsequent experimentation with many methods has led to wide adoption of Johnson's two-stage method of alignment by sorting (2001). This chapter suggests proportional alignment as an improvement to Johnson's method,

Cumpston (2009) suggested a one-stage event alignment method, named "alignment using random selection". It is a one-stage analogue of proportional alignment, and should give similar results. It can be used together with sampling with loaded probabilities, potentially reducing runtimes.

Deterministic tests of four two-stage event alignment methods, including Johnson's method and proportional alignment, are made, using average age at death as a test measure. Stochastic tests of alignment by random selection are made using the Cumpston model. The results are discussed at the end of the chapter. Proportional alignment is a modest improvement on Johnson's method, and alignment using random selection is a one-stage analogue of proportional alignment, giving similar results.

Aligning state numbers can be very difficult, and no standard methods appear to exist.. This chapter describes a three-stage method used to align person ages, areas and types, both in baseline data and after each projection year. The method is complex, and may be fragile in practice. Alternatives to state alignment may sometimes be feasible.

4.1 Alignment needs for household microsimulations

Morrison (2006) gave a history of the evolution of alignment over the three decades in which longitudinal microsimulation had served as a practical policy input. From his paper, there appear to be two broad needs for alignment

- Adjusting assumed probabilities of individual behavior to give overall results consistent with beliefs about the future
- Elimination of stochastic variation, particularly when identifying the consequences of policy changes

Li and O'Donoghue (2011) summarise the main uses for alignment as

- Repairing the unfortunate consequences of insufficient estimation data
- Adjusting for the poor predictive performance of the model
- Producing scenarios based on different assumptions
- Establishing links between microsimulation and macroeconomic models
- Reducing Monte Carlo variability (ie stochastic variation).

Alignment can be used to ensure that projections for past periods match actual experience, and that projections for future periods match projections from other sources.

4.2 Consistency with beliefs about the future

Governments may have national population projections that underpin planning in many areas. Household microsimulations for policy purposes need to have assumptions consistent with those in the national projections, and give comparable results to them.

As Morrison notes, even when considerable care and expertise are used in estimating behavioural equations from past data, very rarely do the microsimulations match expectations about the future.

Harding (2007 p12) asks whether aligning everything at a very disaggregated level may reduce the predictive usefulness of the dynamic model, by imposing upon the micro results predetermined macro outcomes.

4.3 Use of alignment to eliminate stochastic variation

Morrison (2006 p16) comments

> *"In typical policy applications … the focus rests on identifying central tendencies and the numbers of winners and losers resulting from a possible policy change. Clients almost invariably prefer point estimates, and do not usually value information about likely distributions of results."*

As an actuary advising casualty insurers, the author used to encounter a similar attitude. But regulatory authorities for insurers now want information on the range of possible results, including unlikely but costly disasters. Many investors want to understand the risks and potential returns from each investment, as well as the risk reductions from diversification.

The use of alignment to eliminate stochastic variation may thus be concealing valuable information. Note however that stochastic variation during a projection may only be a minor source of overall uncertainty. Uncertainties inherent in the estimation of model coefficients, and changes over time in individual behavior, government policy and economic conditions may be much greater sources of uncertainty.

The US Congressional Budget Office typically makes 500 runs of its CBOLT model, using different assumption sets, so as to quantify the uncertainty inherent in its projections (Simpson 2011).

4.4 Obtaining reasonable projections without alignment

Morrison (2008 p20) noted that DYNACAN had been misinterpreting mortality coefficients inherited from CORSIM, and that alignment had masked the misinterpretation. This appears to be one of the risks of using alignment, where a major underlying error may lead to a variety of distortions. It seems desirable to obtain microsimulation results reasonably comparable with national projections, before applying alignment.

A wide range of problems can cause microsimulation results to differ from national projections. For example, low births may result from low partnering rates or wrong age assumptions about immigrants. Wherever possible, it seems better to correct such problems directly, rather than during alignment.

4.5 Types of alignment

Three broad types of alignment appear to be used

- Event alignment
- State alignment
- Monetary alignment.

Event alignment is typically used to control the numbers of single-outcome events such as deaths. Events with multiple outcomes, such as migration to several alternative destinations, can be broken down into a series of single-outcome events, each aligned separately. This chapter looks at a measure of the performance of five of the many event alignment methods that have been proposed, including two suggested by the author.

State alignment may be needed to correct known biases in baseline data, or to correct deviations from desired state distributions during projections. Examples of both in DYNACAN are provided by Morrison (2008 p14 & 27). Dekkers (2009) describes the extensive use of state alignment by MIDAS-Belgium. Cumpston (2010) describes state alignment procedures used to correct for known biases in the 1% sample file from the Australian 2001 census, and to correct age, area and type errors emerging during projections. APPSIM uses extensive state alignment, for example in cross-sectional alignment of labour force status by sex, age and full-time student status (Harding, Keegan and Kelly 2010 p60).

Monetary alignment tries to ensure that continuous variable such as earnings or investment income match pre-determined totals. It is not examined in this thesis.

4.6 Performance measures for event alignment

Morrison (2006 p15) referred to the “failures of practitioners … to define what constitutes optimality, to assemble a comprehensive set of alignment techniques, and to assess each of these techniques against the appropriate standards”. He suggested the perfect method would “explicitly respect the pool’s *a priori* probabilities, and produce exactly the desired numbers of events, with reasonable correlations of event probabilities across members of the alignment pool” (p9).

Cumpston (2009 p8) used average age at death to assess three death alignment methods. Although only a crude measure of the respect for *a prior* probabilities, this measure was able to differentiate between the three methods, with one having unacceptable performance. Even a crude measure may be enough to rule out some candidates.

Li and O'Donoghue (2011) assess six event alignment methods, using five performance measures and four scenarios based on synthetic data. The performance measures are

- Total deviation index (the difference between simulated and actual events, as a proportion of observations)
- False positive rate (the probability of an event being simulated, given that it does not occur in the data)
- False negative rate (the probability of an event not being simulated, given that it does occur in the data)

- Distribution deviation index (the square of the differences between simulated and expected event numbers, summed over any grouping variable, after weighting by the number of observations in the group)
- Computation time, including restoring the data to their original order.

Total deviation index is not much help, as it was zero for four of the methods tested, and close to zero for the other two. False positive and negative rates were both around 20% for all methods and scenarios, and need to be standardized by deducting the rates for a method considered to be reliable.

The distribution deviation index is a direct measure of respect for the pool's *a priori* probabilities, and discriminated well between methods and scenarios. Two methods, sidewalk hybrid with nonlinear adjustment and Johnson's method, gave very low distribution deviation index values for the scenarios with selection bias and biased alpha intercepts. No methods gave low index values with biased beta coefficients, and it may be asking too much for any alignment method to cope with a wrong assumed probability distribution shape.

4.7 Baekgaard's sampling by sorting for event alignment

In 1995 Baekgaard suggested a "sampling by sorting" alignment method where Monte Carlo simulation is applied to all the eligible persons in the data. Any departure from the alignment total is then corrected by reversing the outcomes of those simulated events where the generated random number is closest to the probability of occurrence (see Baekgaard 2002 p12).

4.8 Johnson's event alignment by sorting

Johnson (2001) proposed "alignment by sorting", a method also involving reversing some of the simulated events. For each prospective event i, a value v_i is calculated, where

$V_i = f(r_i) - f(p_i)$
f(x) has the form -ln((1 - x)/x)
p_i is the prospective event's probability
r_i is a random number drawn from a uniform (0,1) distribution.

The prospective events with the lowest V_i values are then implemented to match the alignment total. In 2006, this was DYNACAN's primary alignment method (Morrison 2006 p6). It is also used in SESIM (Flood, Jansson, Pettersson et al 2005 p15), and in Pensim2 (Redway 2009). It may also be used by LIAM (O'Donoghue, Lennon & Hynes 2009) and MIDAS-Belgium (Dekkers, Buslei, Cozzlino et al 2009).

Table 4.1 Event alignment methods used by national models

Model	Method
APPSIM	Kelly & Percival
Cumpston	Random selection
DYNACAN	Johnson
LIAM	Johnson?
MIDAS-BE	Johnson?
Pensim2	Johnson
SESIM	Johnson

4.9 Kelly and Percival's event alignment method

Under this method (Kelly & Percival 2009 p7-8) every person in scope for an event is assigned a probability score. A large percentage of the desired events are taken from those with the highest probability scores, and the remainder from those with the lowest probability scores. The default percentage in APPSIM is 90%, but this can be user-defined (Harding, Keegan and Kelly 2010 60). If the default percentage is 100%, this appears to be the same method as the sort by predicted probability method (Li and O'Donoghue 2011 p6). Both methods do not use random variables,

4.10 Proportional event alignment

This thesis suggests the use of "proportional alignment", which is similar to Johnson's "alignment by sorting", but uses f(x) = ln(x), rather than f(x) = -ln((1 - x)/x). For x close to zero

$$-\ln((1 - x)/x) \approx \ln(x)$$

so Johnson's method and proportional alignment should give similar results for events with low probabilities.

The intention behind proportional alignment is to keep equal the relative probabilities of different persons being selected for the event, and thus respect the pool's *a priori* probabilities. This objective is achievable, provided the ratio of the numbers of aligned and unaligned events is less than the highest event probability. Note that no alignment method can preserve a priori probabilities if the ratio of the numbers of unaligned to aligned events is lower than the highest event probability.

4.11 Deterministic tests on two-stage event alignment methods

Average ages at death were estimated for the four alignment methods described above. These values were estimated in an Excel spreadsheet, using Australian 2001 mortality rates and 20,000 persons from the 1% sample file from the 2001 Australian census. This file only gives exact ages up to 25, so assumed age distributions were used to impute ages for older persons. The oldest assumed persons were age 104, and the highest assumed mortality rate 0.355. Expected numbers of deaths in a year were estimated as 158.6. Estimates of the effects of the four alignment methods were made assuming aligned deaths of between 0.25 to 4 times expected deaths (ie between 39.6 and 634 deaths).

For Baekgaard's procedure, the expected number of deaths for person i was calculated as

$$p_i + B \qquad \text{subject to a minimum of 0 and a maximum of 1}$$

where p_i is the assumed mortality rate for i's sex and age
and B is an adjustment made for each person to get the desired number of deaths.

For Johnson's procedure, the death of person i is simulated when

$$-\ln((1 - r_i)/r_i) < -\ln((1 - p_i)/p_i) + B$$

where r_i is the random number between 0 and 1 drawn for person i.

Solving for the highest value of r_i resulting in a simulated death gives the expected number of deaths from person i as

$$1 / [1 + 1 / (\exp(\ln(p_i/(1 - p_i)) + B)]$$

For proportional alignment, the death of person i is simulated when

$$\ln(r_i/p_i) < B$$

which gives the expected number of deaths from person i as

$$\exp(\ln(p_i) + B) \qquad \text{subject to a maximum of 1}$$

Excel's Solver was used to find the values of B giving the desired numbers of aligned deaths.

Note that the above analyses are only approximate, as they assume that B is independent of the random variable values. For cases with large numbers of events, this may be a reasonable assumption.

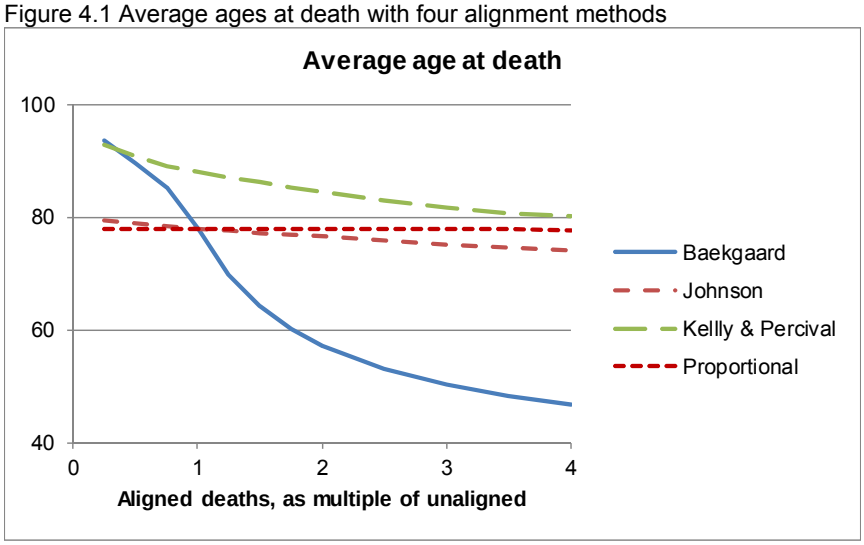
For Kelly and Johnson's method, persons were sorted in order of their probability of death, and the required numbers of deaths selected from the start and end of the list.

For all four methods, all deaths have been treated as a single alignment pool. In practice it is likely that deaths in a number of different age pools would be separately aligned.

4.12 Results of deterministic tests on two-stage alignment methods

Figure 4.1 shows average ages at death estimated using four alignment methods. When the number of aligned deaths was equal to the expected deaths without alignment, the Baekgaard, Johnson and proportional methods gave an average age at death of 77.98 years, the same average as that obtained without alignment.

Figure 4.1 Average ages at death with four alignment methods



Baekgaard's sampling by sorting only works reasonably for alignment totals close to expected. Because alignment is done by reversing the outcomes of those events where the generated random number is closest to the probability of occurrence, most reversals of simulated deaths occur at the young ages. An alignment total below the

expected number of deaths thus leaves most of the old deaths unchanged, and gets rid of many of the younger deaths, giving a high average age on death.

The probability transformation used in Johnson's alignment by sorting means that reversals occur where the difference between the log of the probability and the log of the random number is small. Reversals should thus be shared fairly evenly between young and old deaths, leaving the average age at death fairly stable.

Proportional alignment should keep the expected average at death constant, until such time as the adjusted probabilities become greater than unity. In this case the highest probability was 0.355, so that the average age at death should have remained stable at 77.98 until the ratio of aligned deaths to unaligned deaths became greater than 2.82. The observed average ages were 77.94 with a ratio of 3.0, and 77.49 with a ratio of 4. The calculated results using proportional alignment are also shown in figure 4.2, and compare reasonably well with the results obtained using alignment by random selection.

The apparently poor performance of two of the methods in figure 4.1 may not prevent their use in practice. Choosing alignment pools so as to give reasonably similar probabilities within each band can greatly reduce distortions. Keeping alignment targets reasonably close to unaligned values should be helpful. With Kelly and Percival's method, the default percentage can be chosen so as to give good results when the targets are similar to unaligned values.

4.13 Event alignment using random selection

Cumpston (2009) proposed a one-stage event alignment method, named "alignment using random selection":

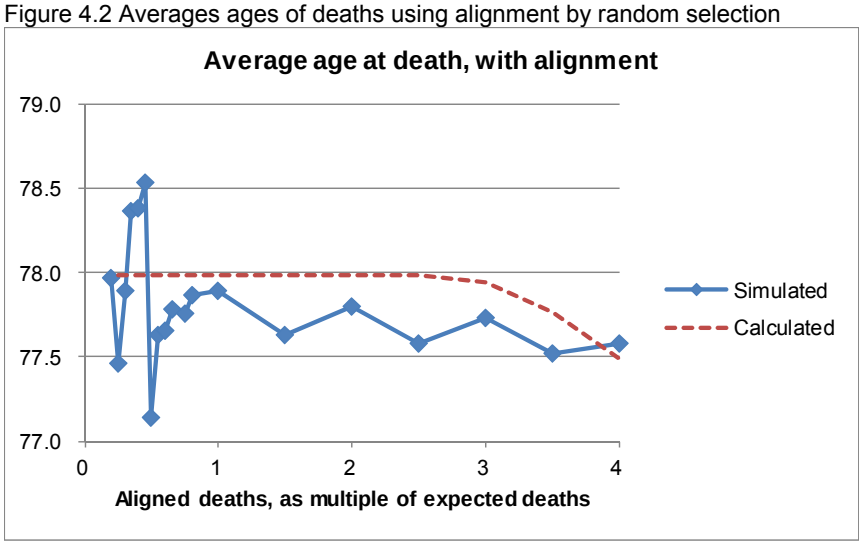
- Obtain the desired event total (for example, the desired total for a month might be one-twelfth of the expected number for the year, for a particular area and age group)
- Randomly select an individual to be tested for the event
- Compare a random number with the event probability for that individual, and decide if the event occurs
- Stop the process when the desired event total is reached.

The intention behind alignment by random selection is to keep equal the relative probabilities of different persons being selected for the event, and thus respect the pool's *a priori* probabilities. It is a one-stage method, expected to give similar results to two-stage proportional alignment. It is similar in concept to the sidewalk shuffle developed by Bouffard and Morrison (Morrison 2006 p14).

4.14 Test simulations of event alignment by random selection

Figure 4.2 shows the average ages at death obtained from one-year simulations of the deaths from 175,000 Australians, using the Cumpston model. Separate alignment totals were calculated for each combination of 8 states and 9 age groups. The age groups used for this purpose were 0-14, 15-24, 25-34 … 75-84, 85+.
Only 1204 deaths were expected in the year, and to smooth out the stochastic variability inherent in numbers of this order, each point on figure 4.2 represents the average of 10 runs.

Figure 4.2 Averages ages of deaths using alignment by random selection



The calculated values in figure 4.2 are the results expected from proportional alignment (also shown in figure 4.1). The calculated values are alignments for Australia as a whole, without subdivision by age or area. Some overall difference in the simulated and calculated values in figure 4.2 is likely, as the source for the calculated values is a 20,000 sample drawn from the 175,000 persons used for the simulations.

The main sources of variability in the simulated results appear to be

- Randomly imputed ages within 5-year age groups for persons over 25 (these are newly imputed for each of the ten trials with each ratio)
- Non-random variations in the share of each of the overall alignment death total between the alignment pools, occurring as the multiple of aligned to expected deaths changes, and the individual pool alignment totals are constrained to integers
- Random variations in the simulations of individuals dying.

Of these sources of variability, the non-random variations in pool shares appear to be the most important, particularly for low numbers of aligned deaths. This problem might be controlled by a more elaborate procedure for choosing individual pool alignment totals. Note that this problem applies regardless of the choice of alignment method.

The simulated results in figure 4.2 were obtained using sampling with loaded probabilities, with maximum probabilities for each pool chosen equal to the highest mortality rate in the pool. To test whether sampling with loaded probabilities was affecting the alignment results, all the tests with aligned deaths higher than expected deaths were repeated, using a maximum probability of 1 for each pool. The resulting average ages at death varied by no more than ±0.02 years from the previous results.

The results in figure 4.2 suggest that

- Event alignment by random selection works over a wide range of ratios of aligned to expected events
- Alignment by random selection gives similar results to proportional alignment.

4.15 Correcting misalignments in baseline data and projections

The baseline data for the Cumpston model is a 1% sample from the 2001 census (ABS 2003d). The ABS does detailed post-census surveys to estimate the extent of census under-reporting, and uses these to publish "estimated resident populations", the starting points for their population projections. There is a tendency for mid-aged persons to be more mobile, and thus less likely to be at home when census collectors call. Ignoring these systematic biases can give significant deviations from national population projections. A related problem is that households with absent members on the census night cannot realistically be used in the base data, further increasing biases from under-reporting.

Three types of state alignment were used in a three-stage process at the start of the projection

- Alignment of the total numbers of persons in each of 9 age groups, done by simulating emigration by persons in over-represented ages, and immigration by those in under-represented ages (892 emigrants and 892 immigrants were needed - see table 4.2)
- Alignment of the total numbers of persons in each age group in each of the 8 areas (done by 1,327 exits of individuals between areas - see table 4.3)
- Alignment of the total numbers of persons of each of 8 person types, for each age group in each area (done by 12,949 exits of individuals within areas - see table 4.4).

A similar process was used at the end of each projection year, to correct age, area and type misalignments. An "exit" is a departure from a household of a person, possibly followed by one or more of the other household members. By contrast, a "move" is where the whole household moves to another dwelling.

4.16 Correcting age misalignments

The first row in table 4.2 shows that 892 emigrants were assumed from over-represented age groups at the start of the projection, balanced by 892 immigrants into under-represented age groups. This is broadly similar to the replication or deletion of small numbers of unattached individuals used by DYNACAN (Morrison 2008 p14).

Subsequent rows in table 4.2 show that some emigrants and immigrants were needed to correct age misalignments after each projection year. When using alignment, the numbers of births, deaths, emigrants and immigrants in each age group were exactly constrained to the numbers of these events derived from Australian national population projections. In theory, the correct number of persons in each age group at the end of the year should have automatically resulted. In practice, it was not possible to exactly replicate national event age groups, and small numbers of emigrants and emigrants were simulated to correct over or under-represented age groups (generally about 400 of each). This is broadly similar to the "illegal immigrants and unrecorded emigrants" used by DYNACAN (Morrison 2008 p27).

Table 4.2 Emigrants and immigrants to align ages

Year	Emigrants to align ages	Immigrants to align ages
0	892	892
1	346	346
2	530	530
3	484	484
4	403	403
5	563	563

4.17 Correcting area misalignments

Movements of households and exits of persons from households were simulated without constraint, using independently derived assumptions. Some of these movements and exits involved changes between alignment areas. At the end of each projection period exits between alignment areas were simulated so as to bring the numbers in each age group into line with national projections. In each case, movements of exit “leaders” were simulated from areas where their age group was in surplus, to areas were in deficit. As exit leaders were sometimes followed by family members of varying ages, an iterative process was needed to reach equilibrium. Given the independent assumptions about movements between areas, it was pleasing that only about 1,000 simulated exits a year were needed to align area numbers (see table 4.3). Given more knowledge of the movement assumptions underlying the national projections, fewer alignment exits might have been needed.

Table 4.3 Exits to align areas

Year	Area misalign-ments	Exits to align areas	Exits per mis-alignment
0	2362	1327	0.56
1	1884	1134	0.60
2	2046	1178	0.58
3	1520	860	0.57
4	1770	1009	0.57
5	1822	1082	0.59

A "misalignment" is a difference between an actual and a target number of persons in any pool, whether positive or negative. On average, about 0.58 exits were needed to correct an area misalignment. This low number is because most of the simulated exits were of persons without followers, and each of these exits corrected a misalignment in the source area and the new area.

4.18 Correcting type misalignments

Numbers of partnership formations and breakups were not available from national projections. Alignments of persons of each type were done against the percentages of persons of each type and age group at the end of each year, which were available from Australian 25-year projections (ABS 2004). Alignments were done by household exits, where exiting persons were assumed to remain in the alignment area (to avoid disturbing previous area alignments). Table 4.4 shows the numbers of exits required.

Table 4.4 Exits to align types

Year	Type misalign-ments	Exits to align types	Exits per mis-Alignment
0	11194	12949	1.16
1	5278	9649	1.83
2	5238	9002	1.72
3	4950	8945	1.81
4	5314	9137	1.72
5	5788	9764	1.69
Average	5314	9299	1.75

Alignment was done in the following sequence

- Persons in non-private dwellings
- Lone parents
- Partners
- Children
- Related persons in family households
- Unrelated persons in family households
- Group households
- Lone persons.

For example, if the number of partners in an area and age group was less than the alignment total, persons of the area and age group, not in non-private dwellings, and not lone parents or partners, were randomly selected, and partners found for them by "best of n" matching. If the numbers of partners were too high, partners of the area and age group were selected, and their departure from partnership simulated. Restrictions were placed on the simulations used to align partner numbers, to avoid altering the numbers of lone parents. Some exits involved multiple persons, and could alter previously aligned totals in other age groups. An iterative process was used to reach equilibrium.

Over the first 5 projection years, about 1.75 exits were needed to correct a type misalignment. No attempt was made to direct persons from one type in excess to another type in deficit, so that something over 1 exit per misalignment was expected.

Authoritative projections of persons by area and type may be hard to obtain, and may not remain valid as experience emerges.

4.19 Complexity of state alignment procedures, and alternatives

The procedures described above for aligning ages, areas and person types are inherently complex, requiring careful programming. They may prove fragile in practice, failing to work because of insufficient persons with particular characteristics.

Commenting on the slow run time for the first version of MIDAS-Belgium, Dekkers (2009) said

> *"we make very extensive use of state alignment … alignment in one state can have unwanted impacts on other states, since the aligned states get 'first choice' in which types of individuals to include. The more complex and intertwined the aligned and non-aligned states become, the larger the room for the wrong individuals to be shifted to the wrong place, or to be moved back and forth. With this … come a large number of ad hoc procedures that are intended to prevent all kinds of 'weird transitions' … (these) gatekeeper and bouncer procedures are relatively slow since they scan through the population vector every time."*

Morand, Toulemon, Pennec, Baggio and Billari (2010 p15-16) note that a better approach is to align by flow of events (such as births, deaths and net migration). Alternatively

> *"Another way to ensure consistency is to use 'calibration' with changing parameter values (ie changing fertility rates) in order that the outputs remain consistent with the macro-level trend."*

4.20 Conclusions

Deterministic tests showed that proportional alignment is a minor improvement on Johnson's method. The other two-stage methods tested had poorer performance, but might still be useful with narrow alignment bands and alignment targets close to unaligned values.

Stochastic testing with Cumpston's model confirmed that alignment using random selection gives similar results to proportional alignment. This one-stage method may provide valuable runtime gains in large models.

In the tests done on alignment using random selection, non-random variations in the proportions of an alignment total allocated to particular pools emerged as an important problem, particularly with low event numbers. This problem results from the choice of alignment pools, and can arise with any alignment method. To minimize this problem, alignment pools should be chosen with large event numbers, and procedures used to ensure that rounding errors do no not cause biases.

State alignment procedures can be complex, fragile and slow. Instead of using state alignment to correct biases in sample data, it may be better to synthesize baseline data from mass data sources, using sample data to ensure appropriate cross-correlations. Instead of using state alignment to ensure projections match expectations exactly, it may be better to change assumptions to provide reasonable closeness.

Calibration, where assumptions are varied so as to give projections closely matching expectations, may be a better alternative to state alignment.

5. MICROSIMULATION WITH WEIGHTED RECORDS

Dekkers has suggested that microsimulation with weighted records is feasible, and may give efficiency gains (Dekkers & Cumpston 2011). The data available for a microsimulation baseline may be weighted records, and it may be helpful to work directly with such records. This chapter describes the method suggested by Dekkers, and gives test results obtained using weighted and unweighted unit records from the Australian 2000-01 Survey of Income and Housing Costs (SIHC) (ABS 2003b).

Each dwelling, whether vacant or occupied, is assumed to have an integer weight. Any persons in the dwelling are assumed to have the same weight as the dwelling. When persons simulated to move into a dwelling have a higher weight than that of the dwelling, copies of the excess persons are assumed to move into other dwellings. Similarly, when the persons have a lower weight, copies of the excess dwelling are created. These copying processes gradually increase the number of persons and dwellings, and reduce the efficiency gains potentially available from weighted microsimulation.

The Cumpston model was modified to store an integer weight for each household, and to do the copying processes needed for moves, exits and immigration. Event alignment during projections of weighted data was done by random selection, using two different methods.

An initial 50-year was made with unweighted data derived by expanding the SHIC weighted data into records each with a weight of one. An unaligned run with SHIC weighted records was made to check that very similar projections were obtained, and two more projections were done with the weighted records, using each alignment strategy. The three tests with weighted records were repeated, assuming no household moves.

Conclusions from these tests are at the end of the chapter. Microsimulation with weighted records works, but the efficiency gains may not justify the additional program complexity.

5.1 Australian data used for testing

Weighted unit records from the Australian 2000-01 SHIC were converted into suitable input files for the Cumpston model. These files covered 16,824 persons, grouped into 6,786 households.

Dwelling weights in the SIHC sample were intended to replicate the Australian population of about 19.4m. To give an unweighted sample size of about 175,000, the weights were multiplied by 0.00937, obtaining dwelling weights ranging from 1.66 to about 34. These weights were then rounded to the nearer integer. The unweighted sample size was chosen so as to be similar to the size of the unweighted HSF file normally used by the Cumpston model. For these tests, the SHIC data were split into 8 regions (the seven states and the Australian Capital Territory), and migration modelled between the 8 regions.

5.2 Model changes to use the weighted data

The Cumpston model was extended to input, store and update a weight for each occupied and vacant dwelling. These code changes needed care, because the microsimulation model treats dwellings separately from households. When a

household moves out of a dwelling, the dwelling is assumed to continue in existence as a vacant dwelling, with the same weight, location, structure, market value and potential rent. The principles suggested by Dekkers were applied:

- A vacant dwelling is selected for a moving household
- if the weight of the moving household is greater than the weight of the selected dwelling, that dwelling is assumed to be become fully occupied, and the weight of the source household is reduced accordingly, with destination and dwelling searching by the source household continuing
- if the weights are equal, the source household occupies the selected dwelling, with the source dwelling becoming vacant, and keeping its original weight
- if the weight of the moving household is less than the weight of the selected dwelling, the selected dwelling is assumed to become occupied with weight equal to that of the source dwelling, and an identical vacant dwelling is assumed in the destination area, with weight equal to the difference between the selected and source dwellings.

Similar principles were used for exits (where an exit is any move of one or more persons out of a dwelling, where at least one person remains in the dwelling), and immigrants. The Cumpston model contained assumptions about the numbers of immigrants each year, and their destination, age and type distributions. The same immigration assumptions were made, with each immigrant household given a weight of 10 (chosen so as to approximately match the average dwelling weight in the SIHC data).

For unweighted data, immigrant families were pre-generated at the start of each projection year. For the weighted data, the numbers of families in the first few years were too low to properly reflect the diversity of immigrants, so immigrant families for each projection decade were pre-generated at the start of the decade, and randomly drawn without replacement, with equal numbers of families being simulated to immigrate in each year of the decade.

Rather than have separate microsimulation models for weighted and unweighted data, the same model was used for both. For unweighted data, all the dwelling weights start at one, and remain at that value.

5.3 Event alignment techniques

Alignment of births, deaths and emigrants was done by random selection. Two different methods were used to deal with weighted records:

1. If the last person selected for an event would cause the alignment total to be exceeded, the household is split, with the excess persons assumed not to have the event.
2. All or none of the persons in the last household are assumed to have the event, depending on which gives the closest approach to the alignment total. Any excess or shortfall from the alignment total for a pool is carried forward to the next projection period.

5.4 Projection assumptions and alignment totals

The SHIC dataset was run with the Cumpston model, using birth, death, emigration and immigration assumptions chosen so as to give a 2051 population estimate approximately equal to the 34.2m projected by the ABS (2008 p39, series B).

Alignment was based on national population projections (ABS 2003c), providing yearly projections of births, deaths, emigrants and immigrants for each region, and subdivisions of these event numbers were obtained for 9 age-groups (0-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+). More recent population projections (ABS 2008) showed higher growth, so the alignment totals were multiplied by 1.383 for births, 1.006 for deaths, 1.077 for emigrants and 1.275 for immigrants. These subdivided numbers were used as alignment totals, for the tests with event alignment. No alignment totals were available for household exits or household moves.

For the weighted data, the numbers of immigrant families in each year were too low to properly reflect the diversity of immigrants, so immigrant families for each projection decade were pre-generated at the start of the decade, and randomly drawn without replacement. Close alignment was obtained each year for births, deaths and emigrants, but alignment of immigrants occurred over each decade.

5.5 Test results

Table 5.1 summarizes key results, which are averages of 5 runs. The table shows both “person records at end” and “persons at end” for each line of results. “Person records” are records of simulated persons, including their household weight, which can be any positive integer. “Persons” are the numbers of simulated persons, obtained by summing the household weights on the person records.

The first line describes the results when the weighting procedure is not applied and the model is run on the ‘expanded’ dataset. The average runtime between 2001 and 2051 was about 26.7 seconds. Applying the weights method proposed in this paper reduced the average runtime by about 33% to about 17.8 seconds. When not allowing for household moves between dwellings, (an assumption implicitly made by models with do not separately simulate households and dwellings), the runtime in seconds is reduced to about 9.2 seconds implying an efficiency gain of about 65%.

Table 5.1 50-year projections using unweighted and weighted data

Weighted Data	Alignment	Moves	Runtime (seconds)	Person records at end	Persons at end	Efficiency gain
No	No	Yes	26.65	312,522	312,522	0.0%
Yes	No	Yes	17.76	310,157	312,689	33.4%
Yes	Method 1	Yes	22.16	305,624	307,806	16.8%
Yes	Method 2	Yes	22.21	305,492	307,806	16.7%
Yes	No	No	9.21	233,813	312,550	
Yes	Method 1	No	12.46	241,139	307,806	
Yes	Method 2	No	11.78	232,817	307,801	

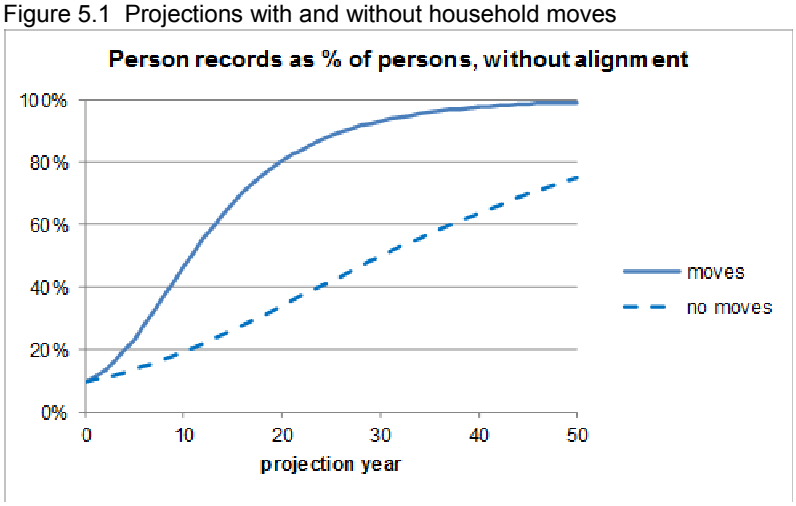
Including alignment methods 1 and 2 in the model increased the average run times. However, even in this case run times remained considerably below the unweighted-no-alignment version of the model. Thus, the efficiency gains remained about 17% and 54% for the model with and without household moves. Because method 2 involves a little less splitting, it should give fewer person records and lower runtimes. The slightly higher runtime for method 2 with moves may reflect random variations in runtimes. Note that the difference between the two methods does not appear to be very important, both in terms of their run times and the alignment totals. The choice of alignment method thus appears not to be crucial in this case.

5.6 Convergence from weighted to unweighted projections

The method presented in this chapter involves the partial expansion or “splitting up” of individual weighted households in case of moves of individuals in between households of different weights, or of households in between dwellings. As a result of a continuous process of moves and the accompanying partial expansions, the average size of the weights will gradually decrease as the dataset becomes more and more expanded. In some future simulation year, all weights will have been reduced to one, and the dataset will have been fully expanded. The total efficiency gain of using weights in the way suggested by Dekkers depends on the average initial size of the weights, and the speed of the convergence process.

As a benchmark, the number of person records of the unweighted (fully expanded) dataset is for all projection years set to 100%. The difference between the curve and the 100% line thus shows the ‘relative expansion’ of the weighted dataset at any moment in time. Figure 5.1 shows that the numbers of weighted person recorded in the unrestricted projections using weighted SIHC data rose quickly, and after 50 years were about 99% of the unweighted numbers. Clearly, the efficiency gain that comes with the approach proposed in this paper is achieved in the first part of the simulation period. The more splitting up is required, the faster the marginal efficiency gain is reduced. The bottom line in Figure 1 reflects the development of the number of person records when moves of households between dwellings are not simulated, and additional splitting is thus avoided. In this case, expansion takes place at a lower rate and the efficiency gain from using weights is therefore larger.

Figure 5.1 Projections with and without household moves



5.7 Conclusions

The method proposed by Dekkers proved robust, and capable of extension to a range of applications. The Cumpston model was able to cope with the additional complications of weighted records, with little increase in memory use or run times. Sampling with loaded probabilities, and event alignment by random selection, proved equally useful with weighted data.

50-year projections with the SIHC weighted unit records gave very similar results to projections with unweighted data derived by replicating each SIHC record in proportion to its dwelling weight.

One alignment method gave exact alignment, but some increase in the numbers of copied households. The other method slightly reduced the numbers of copies, with some minor departures from the alignment totals.
Models which do not model household moves between dwellings will suffer from less record splitting when using weighted records, and should thus have greater efficiency gains.

The test results from Australian data strongly suggest that the method is sound, can be used with or without event alignment, and does give efficiency gains. While the gains were modest with the particular test data, much larger gains might occur in some practical situations. Data with high record weights, and models without moves between dwellings, may offer worthwhile gains.

Using weighted records did however require about 700 extra lines of code, in areas where the model was already complex. Careful coding and debugging were needed, in all taking about 3 weeks. Depending on the availability of programmers familiar with an existing model, adding weighted microsimulation could have significant costs. Microsimulation with weighted records may be useful in special circumstances, rather than generally applicable.

6. MATCHING

Many different types of matches may be needed in household microsimulations. For example, males and females may have to be matched to form partnerships, individuals matched to form groups, and households matched to dwellings. This chapter uses Australian marriage data to test a range of matching techniques.

Past household microsimulation models have simulated marriages on an immediate or batch basis. In immediate matching, a person is selected to be married, and a mate found immediately. In batch matching, all the males and females to be married are identified, and then paired off.

Immediate “best of n” and probability-weighted matching methods, suggested by Cumpston (2009a, 2010a), are amongst the methods tested. Immediate probability-weighting uses the same probability weighting technique as stochastic matching, a widely used batch matching method.

Logistic regression is sometimes used to derive compatibility measures for prospective marriage partners. For the Australian marriage data used here, logistic regression is shown to be a little less precise than least squares fitting, but more versatile.

A number of performance measures are here used to test the fidelity of matching methods. Sums of squares of fitting errors, chi square statistics, and standard deviations of percentile fitted/actual ratios, are all singe-figure measures. Graphs of actual and expected age differences, and of the standard deviations of these differences, can help decide whether fitting errors are likely to matter in practice.

The test results are discussed at the end of the chapter. Immediate matching methods seem likely to give better fidelity than batch matching, be valid with small and large regions, and give lower runtimes.

6.1 Early matching processes

Orcutt, Greenberger, Korbel and Rivlin (1961 p93) selected individuals to be married, and then used functions describing the relations between persons marrying to randomly select the characteristics of the partner. Individuals to be married were held in temporary storage until a person with the selected characteristics was identified during the pass through the household file.

Orcutt, Caldwell and Wertheimer (1976 p34-35) selected males and females to be married during a year, then ranked them based on race, age, education and region. After the rank orderings of males and females were merged, the excess male or females who happened to fit least well were returned to the single population.

These two early processes have many of the features of matching processes used subsequently. In the model described in 1961, marriage partners were found as soon as possible for persons selected to be married. In the model described in 1976, males and females to be married were selected during the year, then merged in a batch process at the end of the year, with the poorest fits being rejected.

Bouffard, Easther, Johnson, Morrison and Vink (2001 p6) noted that the CORSIM, DYNACAN, POLSIM and SVERIGE models all used something similar to CORSIM’s original “stable marriage algorithm” to form specific couples from pools of prospective marriage partners. In spite of its theoretical elegance, this was found to produce too many close partnerships, and too many distant ones.

6.2 Matching methods used in recent national models

Table 6.1 gives some examples of the matching methods used in recent models.

Table 6.1 Matching methods used in recent national models

Model	Version	Immediate or batch	Matching method
APPSIM	2009	Immediate	Statistical selection from eligible partners
CBOLT	2002?	Batch	Normalised match probabilities
Cumpston	2009	Immediate	Best of n
DYNACAN	2002?	Batch	Stochastic
DYNASIM3	2002?	Batch	Random within race, age & education
INAHSIM	1986	Batch	Merge by age
MIDAS-BE	2008	Batch	Order of decreasing difficulty
Pensim2	2004	Batch	Order of decreasing difficulty
SESIM	1997	Immediate	First male 3 years older in region
SMILE	2005?	Batch	Order of decreasing difficulty
SVERIGE	2002	Batch	Replication of patterns of recent pairs

Sources
APPSIM - Bacon & Pennec (2007 p19)
CBOLT - Peres (2002 p17)
Cumpston - Cumpston (2009 p13-19)
DYNACAN - Leblanc, Morrison & Redway (2009 p11-12)
DYNASIM3 - Favreault & Smith (2004 p9)
INAHSIM - Inagaki (2009a p6)
MIDAS-BE - Dekkers et al (2009 p33)
Pensim2 - Emmerson, Reed & Shephard (2004 p30)
SESIM - Pettersson 2009
SMILE - O'Donoghue, Lennon & Hynes (2009 p24)
SVERIGE - Holm, Holme, Makilla, Maffsson-Kauppi & Mortvik (2004 p20-23)

6.3 Batch matching by stochastic matching

Stochastic matching, proposed by Easther and Vink in 2000, calculates the relative probability of each potential pairing from two equal pools of males and females, and randomly selects a pairing on a probability-weighted basis. This proceeds sequentially until pairing is complete. This process was implemented in CORSIM and DYNACAN.

Easther & Vink (2000) showed analytically that two versions of stochastic matching, when applied to preselected equal batches of males and females, produce marriage probabilities that are proportional to the compatibility measures.

Leblanc, Morrison and Redway (2009) tested stochastic matching, tournament selection and order of decreasing difficulty matching, and concluded that none gave good results, but stochastic matching was clearly the best of the available methods.

6.4 Batch matching in CBOLT, INHASIM and SESIM

Perese (2002 p17) proposed a sequential matching process for male and female pools, where the probabilities of marriage to a male are calculated for each remaining female, and normalized so that the highest probability is 1. For each female in turn, a random number is drawn, and a match made if the number is less than the calculated

probability. As a result of the normalization, a match is always found. This process was implemented in CBOLT, apparently on a batch basis.

INAHSIM selects equal numbers of males and females to be married, sorts both by age, them merges (Inagaki 2009b). A randomised version of this process is tested later in this chapter.

SESIM randomly selects females in a region to be married, and creates a pool of all eligible males in the region. For each selected female in turn, a male 3 years older is selected. If no such male can be found, age differences of 2 to 4 years are considered (Pettersson 2009).

6.5 Batch matching by order of decreasing difficulty (ODD)

Leblanc, Morrison and Redway (2009) tested a modified version of a matching procedure proposed by Redway. Good matches are first made for those hardest to pair off, with pairing proceeding in order of decreasing difficulty. Although this algorithm did extremely well in terms of not creating marriages with extreme age differences, it also generated far too many marriages in which the husband is a single year of age greater than his wife. The authors suggested that the poor quality of the compatibility measure used prevented any true testing of matching procedures. Matching by order of decreasing difficulty is being used by PENSIM2, and by LIAM-based models such as SMILE (O'Donoghue, Lennon and Hynes 2009) and MIDAS-Belgium (Dekkers et al 2009). A version tested later in this chapter gave poor fits to Australian marriage data.

6.6 Batch matching by tournament selection

Leblanc, Morrison and Redway also tested a version of tournament selection, where a number of potential partners for a person are randomly drawn from the pool, and the best of these selected. The number of persons drawn is taken as 10% of the total pool, with some defined minimum number. The authors tested minimums of 5, 10 and 20 persons. In each case, they found that the procedure generated far too many marriages with extreme age differences. When used with large pools, this process seems likely to give too many close matches.

6.7 Immediate "best of n" matching

Cumpston (2009a) examined the use of "best of n" matching, where a male or female is simulated to marry, n potential matches for the person are randomly selected, and the match with the highest compatibility score selected. The limited number of selections can make this matching method faster than other methods, but it needs careful tuning to give reasonable matches.

6.8 Immediate probability-weighted matching

Cumpston (2010a) suggested the use of immediate probability-weighted matching. As soon as a male or female is simulated to marry, some or all of the potential marriage partners for that person are identified. One of these potential partners is then randomly selected, with a procedure where a partner's probability of selection is proportional to their probability of marriage to the person.

Although this procedure appears similar to stochastic matching, there are important differences. Probabilities have to be those for marriage to randomly selected single persons, rather than to persons already selected to marry. The numbers of potential matches are much higher, so that calculation times can be much longer than for stochastic matching. As batches are not used, there is no problem with the last matches being unlikely.

The calculation times needed for immediate probability-weighted matching can be greatly reduced if selections are made from a small number of randomly sampled single persons, rather than from all single persons.

6.9 Performance tests for different matching methods

A variety of criteria have been used to test the validity of matching methods. If these criteria are too weak, then quite poor matching methods may be accepted as useful. As figure 6.1 shows, age differences and age standard deviations for persons marrying are strongly dependent on the age of the male, and comparisons of overall age differences may fail to detect major problems at particular ages.

Bouffard, Easther, Johnson, Morrison and Vink (2001) used graphs of the percentages of couples with each age difference, as well as percentile graphs of compatibility levels. These graphs allow visual comparisons between the pairing results and population values. The authors noted (p4)

> *"…the synthetic marriages should resemble actual new marriages in both their central tendencies and their dispersion"*

This chapter uses chi square test statistics, calculated for each age combination for which at least five marriages are observed in the data. These statistics are commonly used for goodness of fit tests, and are used for that purpose here. Particularly for large numbers of observations, chi square goodness of it tests may fail, even though visual examination suggests that a reasonable fit has been obtained. For this reason, a number of graphs have been included, showing observed and fitted age distributions of females marrying males of each age. Graphs of means and standard deviations of age differences at marriage also help assess the extent to which the synthetic marriages resemble actual new marriages.

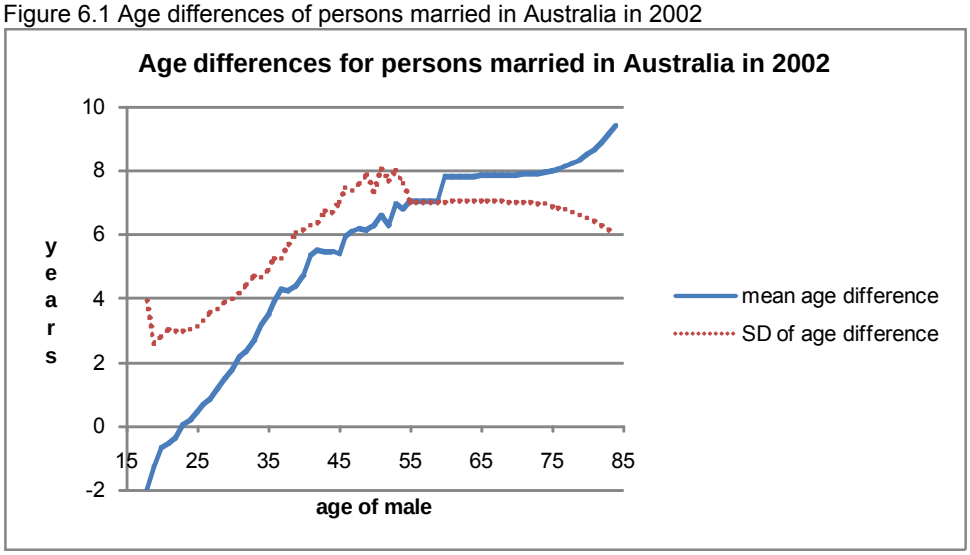
6.10 Persons married in Australia in 2002

Test data were derived from data on the ages of persons in 105,435 new marriages in 2002 (ABS 2003e). The data show the cross-classified ages of brides and grooms in single-year age groups up to 54, then 55-59, 60-64, and 65 and over. In addition, the ages of males marrying are shown in single years up to age 74, then 75 and over, as are the ages of females marrying. The data showed very few persons marrying below age 18, and these persons were omitted from analysis. To fill the data gaps over age 54, it was assumed that, for each age of a male at marriage, the ages of brides were normally distributed. Normal distributions for each male age were obtained by minimising the sum of the squares of differences of the actual and fitted numbers for each single age or age-group.

The mean age differences in figure 6.1, and their standard deviations, were obtained from the actual one-year data where available, and from the fitted normal curves where one-year data were not available. The mean age differences show that very young males tend to marry females a little older than themselves. After age 23 males

become progressively older than their brides, with the estimated age difference being 6.8 years for males aged 54. The fitted standard deviation of female ages for males aged 18 is 3.9 years, dropping to 2.55 years at age 19, and then climbing to 7.6 years at age 54. The low numbers of persons marrying after age 54, and the unavailability of single-year data, create very wide confidence limits for the estimates for males over 54. The mean age difference and its standard deviation may in fact continue to increase with male ages well past 54.

Figure 6.1 Age differences of persons married in Australia in 2002



Orcutt, Greenberger, Korbel and Rivlin (1961 p90-91) gave probabilities of males in each 5-year age group marrying females in each age group, separately for never married and remarried. Other variables, such difference in years of education, number of children, race, labour force participation and earnings difference have also been considered relevant to the probabilities of individuals marrying each other.
While this Australian data only contains ages, it is sufficiently detailed to test fitting methods, and to identify matching methods unlikely to work with more complex data. Its simplicity made it possible to write fast algorithms to exhaustively test different matching methods.

6.11 Alternative models and fitting methods

Table 6.2 compares four different models fitted to Australian 2002 marriages using least squares regression. The best of these models, using piecewise quadratic functions, was also fitted using logistic regression.

Table 6.2 Performance measures for models fitted to Australian marriage data

Probability model	Fitting method	Degrees freedom	Sum of squares	Chi square
Quadratic	Least squares	1,309	672,472	4,880
Gamma	Least squares	1,242	1,576,369	8,323
Quartic	Least squares	1,175	727,746	4,397
Piecewise quadratic	Least squares	1,237	122,323	1,233
Piecewise quadratic	Logistic	1,237	214,934	1,468

For the quadratic model, the relative probability of a male age x marrying a female age y was defined as

$$\exp(\text{score}) / (1 + \exp(\text{score})) \quad (6.1)$$

where $\text{score} = a + b(x - y) + c(x - y)^2$ (6.2)
x is the age of the male
y is the age of the female
and a, b and c are regression coefficients to be fitted.

The regression constant a was held at -3 throughout, as this constant makes little difference to the use of relative probabilities for matching. Regression coefficients b and c were fitted for each age of male by minimizing the sums of squares of the differences between the numbers of marriages with each combination of male and female ages. The observed numbers were the single-year data published for 2002, plus single-year estimates derived from the normal curves fitted as described in 6.10. Fitting was done within a spreadsheet, using Excel's Solver, fitting each year of age for males separately. Degrees of freedom were calculated from the 1,443 cells with at least 5 observed marriages, less the 134 fitted parameters. The chi square test statistic was calculated as

$$\sum\sum (O_{ij} - E_{ij})^2 / O_{ij}$$

where E_{ij} is the number of females age j expected to marry males of age i, as estimated from the fitted coefficients and equations
and O_{ij} is the number of females age j who did marry males of age i in 2002, estimated as described in 6.10.

The chi square statistic was calculated by summing across all cells with at least 5 observed marriages. Observed marriages were used in the denominator, as this gave more stability when comparing different fits. As the mean of a chi square distribution with k degrees of freedom is k, and its variance is 2k, a chi square value of 4,880 with 1,309 degrees of freedom reflects a poor fit. Trials with gamma and quartic relative probability functions gave similar or worse fits, both on a sum of squares basis and on chi square statistics. These poor fits may have been partly due to the difficulties encountered in fitting these more complex functions.

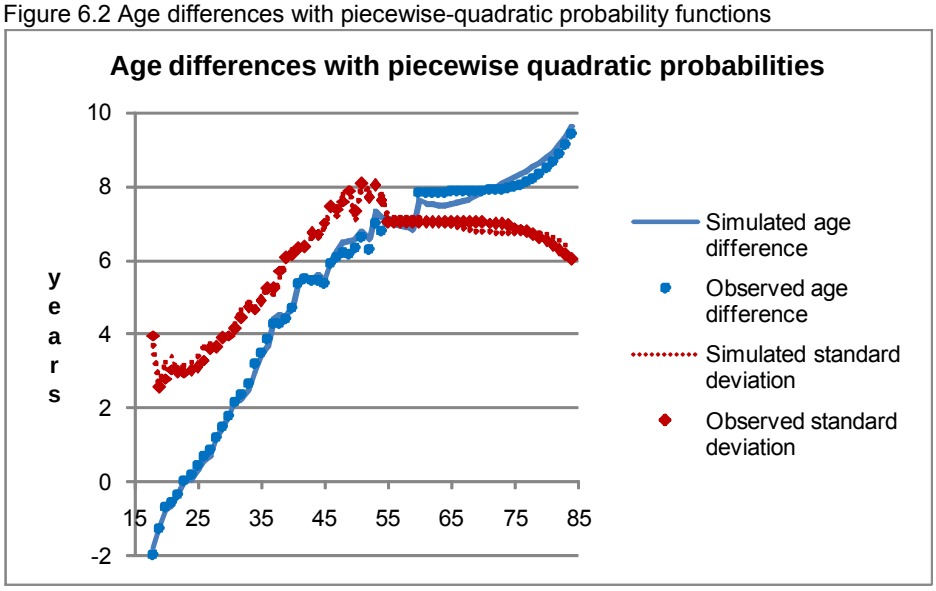
Better results were obtained by fitting piecewise-quadratic functions for males aged 19 to 54, where a function such as in equation 6.2 was used for all values of female ages below the male age, and the same function but with different coefficients used for all female ages equal to and above the male age. From age 55 on, a single quadratic function proved adequate at each male age. The chi square statistic of 1,233 is close to the expected value of 1,237.

For ages 18 to 36, the fitted coefficients of the age difference square for female ages above the male age were positive, giving rising probabilities for females many years older than males. To avoid unrealistically high probabilities of marriage to much older females, the age difference in the last term of equation 6.2 was assumed to be not more than 12 years, for those cases where the female was older than the male.

6.12 Age differences obtained with piecewise-quadratic relative probability functions

With a few exceptions, the fitted and observed mean age differences, and their standard deviations, agree reasonably closely (see figure 6.2). This is true even at the older ages, where the data are thin and partly artificial. For practical purposes, piecewise-quadratic relative probability functions appear to be giving reasonable fits for males up to age 54, and quadratic functions good fits from 55 on.

Figure 6.2 Age differences with piecewise-quadratic probability functions



6.13 Fitting relative probabilities of marriage using logistic regression

CORSIM and DYNACAN used compatibility measures derived from logistic regression using census data. Variables used included age difference and its square, difference in years of education, number of children the women has, race, labour force participation, earnings difference and some interaction effects (Bouffard et al 2001 p5). The compatibility index was estimated using logistic regression on a potential pairs data file of recent marriages. The potential pairs may have included each possible pairing of the persons recently married, as well as the pairs which did marry (Perese 2002 p9).

Logistic regression is widely used to estimate probabilities of events occurring, and the data needs to have some cases where the event occurs, and others where it does not. The addition of potential pairs is thus a device to make the regression feasible, and there is no need to add any particular number of potential pairs.

Logistic regression fits probability models of the form

$$\text{score} = \beta_0 + \sum \beta_i x_i \qquad (6.3)$$

$$\text{probability} = \exp(\text{score}) / (1 + \exp(\text{score})) \qquad (6.4)$$

where β_i are the fitted parameters, and x_i the explanatory variables.

If the potential partner with the highest probability is to be selected, then comparisons can be made between regression scores, and the value of the constant β_0 is

immaterial. But if partners are to be selected on a probability-weighted basis, as in stochastic matching, then the value of β_0 will matter. The number of potential pairs added for the regression analysis will thus have some effect on the results of stochastic matching.

6.14 Logistic regression applied to Australian marriage data

To test the validity of using logistic regression with dummy pairs, the technique was applied to the Australian 2002 marriage data described in 6.10. 105,372 records were created showing the ages of males and females marrying, using the exact ages up to age 54, and the fitted ages above 54. In addition, dummy records were created with the same male ages as the males known to have married, and female ages obtained by random sampling from the 3,347,851 single females aged 18 to 84 in Australia in 2001. Three sets of dummy non-marriage records were created, with 1, 3 and 10 dummy records for each of the 105,372 records representing actual marriages. The fitted logistic models were of the form described in 6.13.

Some of the observations made from these trials were

- All three data sets gave large log-likelihoods at each male age, but the magnitudes of the reductions reduced with the numbers of dummy records (by about 0.1 for a tenfold increase in dummy records at typical marriage ages)
- The fitted regression constants reduced in magnitude with the numbers of dummy records, with the reduction being approximately equal to the natural logarithm of the ratio of the numbers of dummy records
- Very similar regression coefficients were obtained at each age, with some variations at male ages of 70+, and with some variations in the quadratic coefficients at peak marriage ages
- Estimation errors for the regression coefficients were largely independent of the numbers of dummy records for all ages up to about 70, and were then strongly dependent on the numbers of dummy records, with more records giving lower estimation errors
- Given these results, as many dummy records as feasible should be used
- No convergence problems were observed.

6.15 Results using logistic regression to fit probabilities

The first line of table 6.3 is as in table 6.2, obtained by using Solver to fit piecewise-quadratic relative probabilities so as to minimize sums of squares. The second line gives the results obtained by fitting piecewise-quadratic probabilities using logistic regression on data created as in 6.14, with 10 dummy records per actual marriage. Sums of squares, and the chi square statistic, are both higher with probabilities obtained by regression. Logistic regression maximizes the likelihood of the observed results, and will generally not give the same results as minimizing sums of squares.

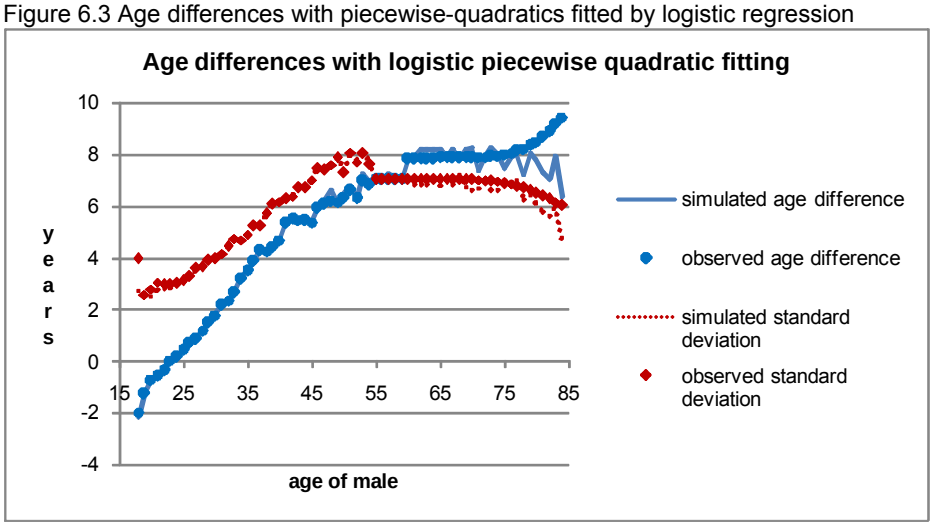
Table 6.3 Results with piecewise-quadratic probability functions

Derivation	Degrees freedom	Sum of squares	Chi square
Solver	1,237	122,323	1,233
Regression	1,238	214,934	1,468

Comparing figures 6.2 and 6.3 suggests that piecewise-quadratic probability functions obtained by logistic regression with dummy records are almost as good at replicating means and standard deviations of age differences as functions obtained using

spreadsheet analysis. This is valuable, as logistic regression is easier to use, and provides information about estimation errors and significance.

Figure 6.3 Age differences with piecewise-quadratics fitted by logistic regression



6.16 Method comparisons using logistic compatibility measures

The piecewise quadratic models obtained with logistic regression were used in table 6.4 to compare a number of proposed and existing matching methods.

Table 6.4 Results with different methods using logistic regression models

Immediate/ Batch	Matching method	Sample size	Sum of squares	Chi square	Percentile SD
Immediate	Probability weighted	3,347,851	214,934	1,468	0.044
Immediate	Probability weighted	200	259,819	2,700	
Immediate	Probability weighted	100	279,732	2,709	
Immediate	Probability weighted	50	308,535	2,816	
Immediate	Probability weighted	30	364,242	3,049	0.068
Immediate	Probability weighted	20	497,016	3,457	
Immediate	Probability weighted	10	1,026,008	5,476	
Immediate	Best of n	4-6	484,912	3,879	0.090
Batch	Stochastic	105,372	409,197	5,636	0.078
Batch	Age ordering, with randomisation	105,372	1,062,470	6,163	0.139
Batch	Order of decreasing difficulty	105,372	185,184,957	621,141	1.485

Each test in table 6.4 simulated the marriage of 105,372 males with the age distribution of males actually married in 2002. The first test of probability-weighted matching in the table was made by selection from each of 3,347,851 single females aged 18 to 84. The other tests of immediate methods were made assuming selections of partners from samples of 200 or less. The tests of batch methods were made selecting from 105,372 females with the age distribution of females actually married in 2002. Percentile standard deviations were derived as described in the next section.

6.17 Immediate probability-weighted matching from samples of 30

Table 6.4 shows that immediate probability-weighted matching with samples of 30 gave inferior test-statistics to the same method applied to all eligible partners. For example, the Chi square statistic deteriorated from 1,468 to 3,049. Figure 6.4, however, shows that means and standard deviations of the age differences for the simulated marriages were very close to values for actual marriages in 2002.

Figure 6.4 Age differences with probability-weighted matching from samples of 30

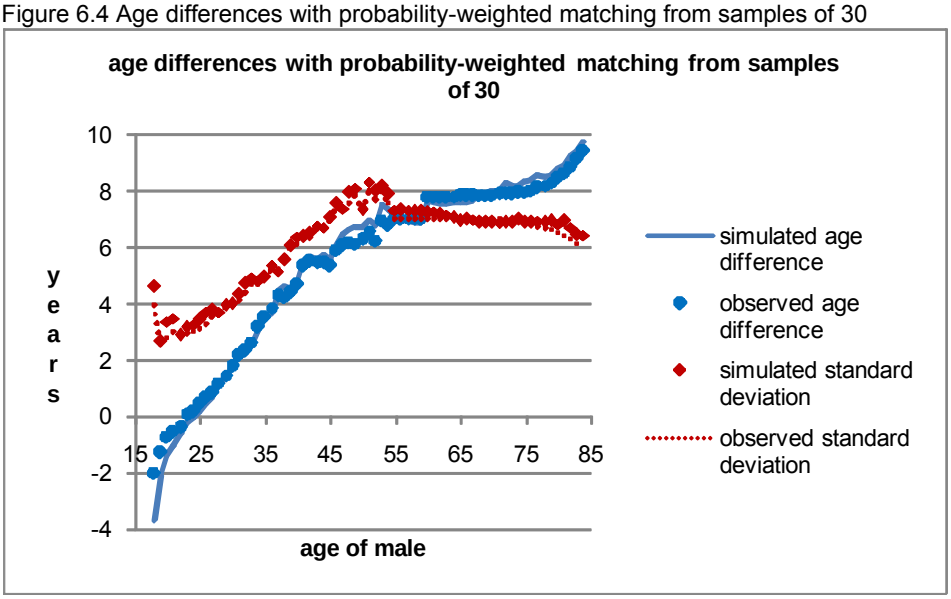
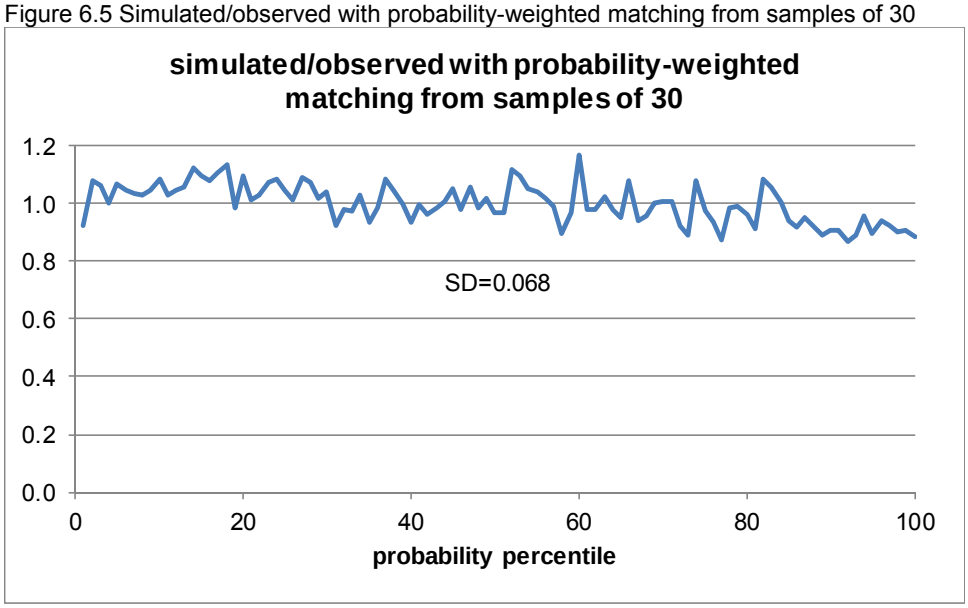


Figure 6.5 shows the numbers of fitted marriages, divided by the numbers of observed marriages, for each probability percentile. These percentiles were derived by taking each male/female age combination, calculating the relative probability of marriage, and sorting by probability. Numbers of observed marriages were then added, with each age combination being assigned to the probability percentile applying to its last member. This process resulted in probability percentiles averaging 1% of marriages, but with some variations around this size.

Figure 6.5 Simulated/observed with probability-weighted matching from samples of 30



The simulated to observed ratios in figure 6.5 have an average of 1.0, and a standard deviation of 0.068. There appears to be a slight bias for low probability marriages, and against high probability marriages. Probability-weighted matching from samples of 30 appears to be working reasonably well in this instance.

While samples of about 30 appear to give reasonable results with the simple test data used here, larger samples may be needed in practice. Different sample sizes may be appropriate for persons of different ages.

6.18 Immediate matching with "best of n" selection

Another way to reduce the numbers of potential partners tested is "best of n" matching, where the best of n randomly selected partners is chosen. This was suggested by Cumpston (2009a), and tested using Australian partnership data. Taking the best of 8 random selections was there found to reasonably reproduce the overall mean and standard deviation of the age differences between existing partners. Best of n matching is tested here using piecewise-quadratic probability functions, and found to give reasonable results with 4, 5 or 6 samples, depending on the age of the person simulated to marry.

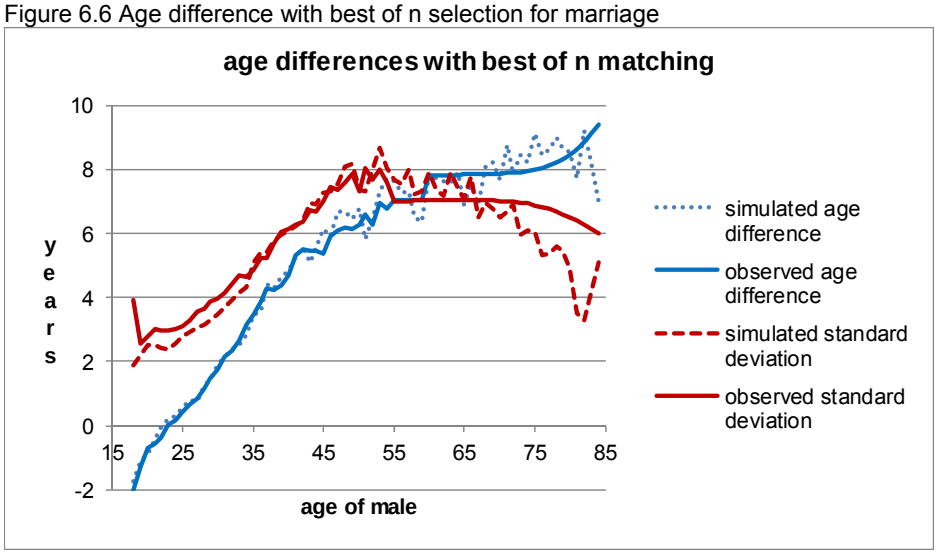
Table 6.5 shows the numbers of samples found to give reasonably low chi square statistics when using "best of n" sampling from single females aged 18 to 84 in Australia in 2001. 100 trials were made for each of 105,372 males assumed to marry, using piecewise-quadratic probability functions. Potential brides were restricted to females not younger than the male by more than 35 years, and not older than the male by 25 years. The numbers of samples in table 6.5 were in general chosen as those giving the lowest chi square statistics, with a few exceptions made to give simple age ranges.

Table 6.5 Number of samples

Age of male	Samples
18-19	6
20-34	5
35-49	4
50+	5

Best of n gives test statistics that are a little worse those obtained with probability-weighted matching from samples of 30 (see table 6.4).

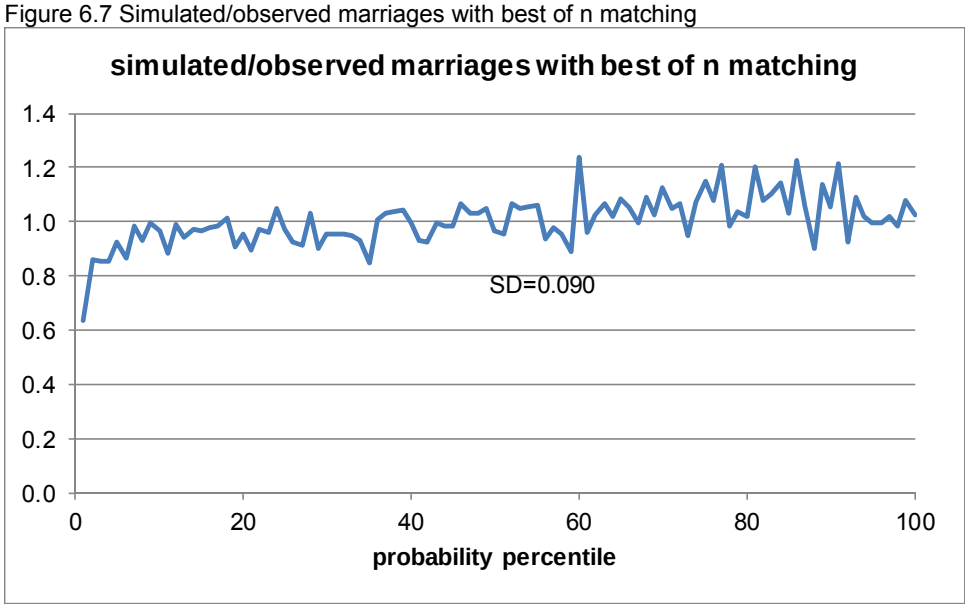
Figure 6.6 Age difference with best of n selection for marriage



The simulated standard deviations are noticeably further from the observed values than those in figure 6.4, particularly for age 55 on. Given that about 95% of marriages involve males below 55, these results suggest that “best of n” selection, appropriately used, can reasonably implement probability-weighted matching for marriages

The simulated to observed ratios in figure 6.7 have a standard deviation of 0.09. There is a slight bias against low probability marriages, and for high probability marriages.

Figure 6.7 Simulated/observed marriages with best of n matching



6.19 Stochastic matching

Unlike probability-weighted and best of n matching, stochastic matching involves selecting all the males and females to be married in a period, then pairing them off:

- using probabilities of marriage, equal numbers of single males and females are randomly selected as persons to be married
- the probability of each selected male being married to each selected female is calculated
- the resulting probabilities are added across all the potential marriages, and a single one of these randomly selected on a probability-weighted basis
- the married male and female are removed from the pool, and steps 3 and 4 iterated until all the males and females have been married.

In this instance, the ages of 105,372 males and females aged 18 to 84 who married each other in 2002 were available. Probabilities of marriage between these persons were required, rather than probabilities of marriage with single persons. For each record of a marriage, a further ten non-marriage records were created with the same age as the male, but with female ages randomly chosen from those of females marrying in 2002. Logistic regression was used to fit relative marriage probabilities for the following groups

- a quadratic function for males aged 18
- piecewise-quadratic functions for each year of male age from 19 to 29
- quadratic functions for each year of male age from 30 to 59
- a quadratic function for males aged 60+.

As stochastic matching requires absolute rather than relative probabilities, the derived regression constants were used, rather than replaced by a common value. Initial trials showed that stochastic matching with the fitted functions resulted in some very old males being left unmarried till near the end of matching, causing some very improbable matches between old males and young females. As a rough fix for this problem, the regression constants for male aged 18 to 59 were reduced by 4. Stochastic matching is inherently more complex than probability-weighted matching, and it is not clear how matching probabilities should be determined.

Table 6.4 shows that the goodness of fit measures for stochastic matching were a little worse than for probability-weighted matching with samples of 30. Figure 6.8 shows that stochastic matching gave simulated age differences that were a few years higher than observed from about age 50 on, and standard deviations that were appreciably higher at some years. A more careful process for deriving probabilities of marriage between particular males and females might have given better results.

Figure 6.8 Age differences obtained with stochastic matching

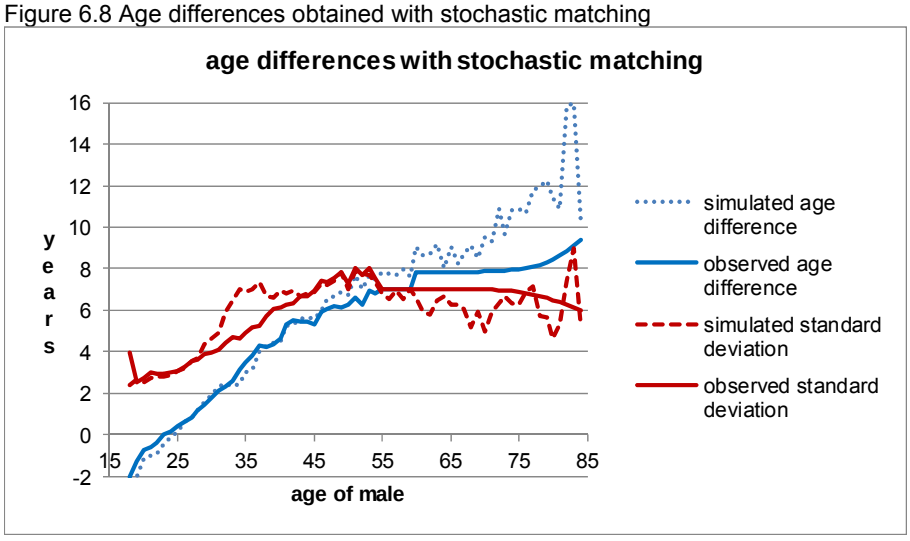
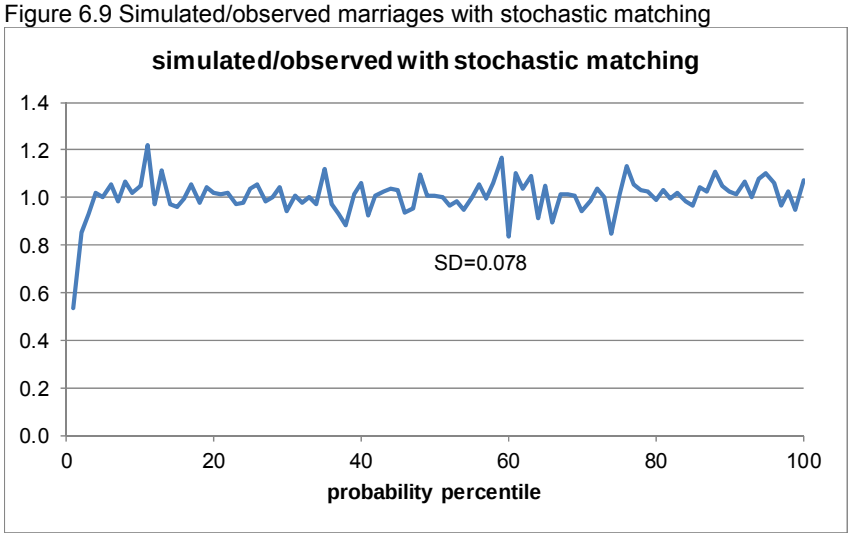


Figure 6.9 shows show percentile fits with a standard deviation of 0.078, and little evidence of any systematic bias toward high or low probability marriages.

Figure 6.9 Simulated/observed marriages with stochastic matching



6.20 Batch matching by ordered merging

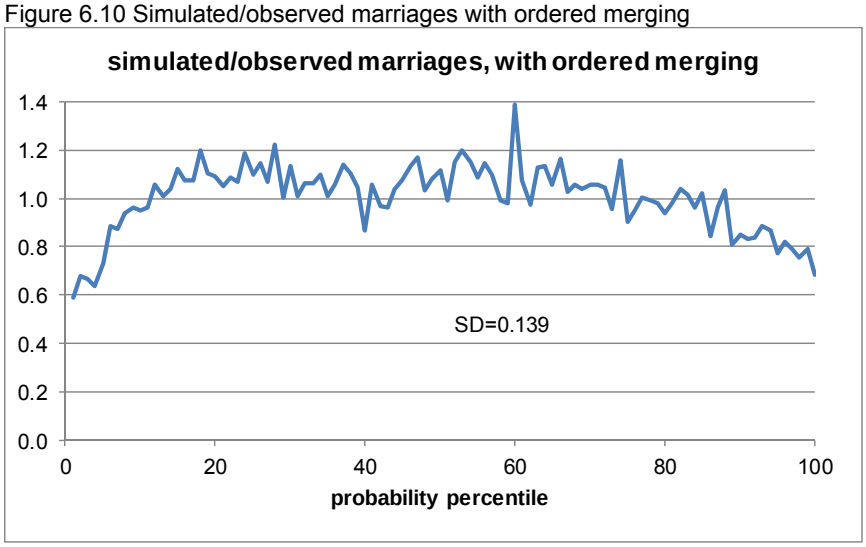
The INAHSIM model simulates marriages on a batch basis by sorting the selected males and females by age, then merging. Tests were made here with a randomised version of this procedure:

- A randomised age was calculated for each of the 105,372 males married in Australia in 2002, calculated as their actual age plus a random variable simulated from a normal distribution with zero mean and standard deviation equal to 1.2 times the estimated standard deviation in ages of females marrying males of that age
- The males were sorted in the order of their randomised ages, and merged with the females sorted in the order of their actual ages.

The 1.2 multiple of the standard deviation was chosen so as to give a low standard deviation for the simulated/observed marriages. A multiple of 1.0 gave a standard deviation of 0.210, 1.2 gave 0.139 and 1.4 gave 0.153.

Figure 6.10 shows that too few marriages with low probabilities, and too few with high probabilities, are being simulated. The test statistics in table 6.4 are worse than for stochastic matching, but a good deal better than those for ODD matching.

Figure 6.10 Simulated/observed marriages with ordered merging



Ordered matching is a simple robust procedure, but it is difficult to see how it can be generalized to take into account variables other than age.

6.21 Order of decreasing difficulty (ODD) matching

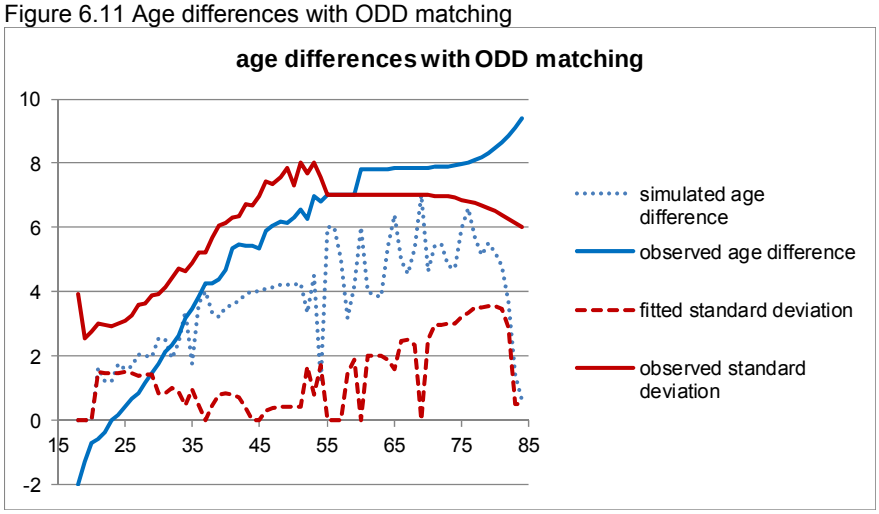
The version of ODD matching here tested is based on the version described by Leblanc, Morrison & Redway (2009 p15):

- The central male is chosen to have the average age for all males marrying in 2002, and the central female to have the average age for all females marrying
- Select the person furthest in age from their central age (whether male or female)
- From all the persons of the opposite sex to the selected person, find the person with the highest probability of marriage to the selected person
- Marry these two persons, and iterate the process until everyone is married.

Marriage probabilities were taken to be those derived for stochastic matching, except that all regression constants were assumed to be -3.

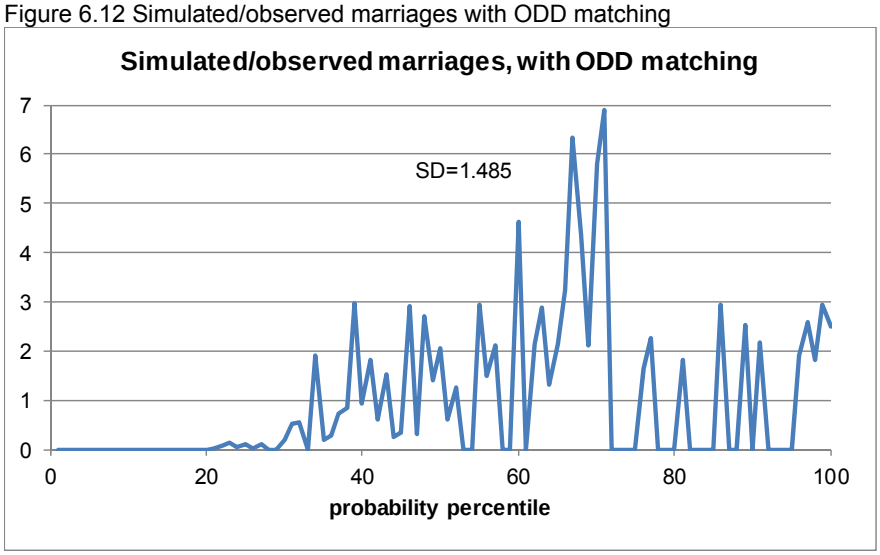
Figure 6.11 shows that average age differences were too low for males aged 37 and over, while standard deviations of age differences were much too low at all ages.

Figure 6.11 Age differences with ODD matching



As figure 6.12 shows, ODD matching gave too many moderate and high probability marriages, with almost no low probability marriages. These results are compatible with the observation by Leblanc, Morrison and Redway (2009) that ODD does extremely well in terms of not creating marriages with extreme age differences, but also generates far too many marriages in which the husband is a single year of age greater than his wife. Figure 6.12 confirms this poor performance.

Figure 6.12 Simulated/observed marriages with ODD matching



Trails with a modified version of ODD using probability weighting (Cumpston 2010 41) show that this greatly improves the performance of the method, but it is still inferior to stochastic matching.

6.22 Numbers of compatibility calculations

Compatibility calculations using logistic models can use appreciable computer time. Table 6.6 shows the approximate numbers of compatibility calculations with different methods and numbers of regions.

Table 6.6 Numbers of compatibility calculations with different methods

Persons m	Regions	Marriages per region	Method	Sample size	Compatibility calculations m
2	1	10000	Probability weighted	30	0.3
2	1	10000	Best of n	5	0.05
2	1	10000	Stochastic	10000	100
2	1	10000	ODD	10000	50
2	50	200	Probability weighted	30	0.3
2	50	200	Best of n	5	0.05
2	50	200	Stochastic	200	2
2	50	200	ODD	200	1

Each line of table 6.6 assumes that there are 2m persons in the model, and that there are likely to be about 10,000 marriages a year. Immediate probability-weighted matching with samples of 30 will need 30 compatibility calculations for each of the 10,000 marriages per year, or about 0.3m calculations per year. This figure will be unchanged if marriages are separately modelled for each of 50 regions. Best of n sampling will similarly require about 5 calculations per marriage, or about 0.05m calculations, regardless of the use of regions.

Stochastic matching between 10,000 males and 10,000 females selected for marriage involves initial calculation of compatibility measures between each male and female, or 100m calculations. If there are 50 equal sized regions, then compatibility measures have to be calculated between each of 200 males and 200 females, giving 40,000 calculations per region, or 2m calculations in all.

ODD matching involves requires the same number of calculations as stochastic matching for the first pair matched, and then linearly reducing numbers of calculations as the numbers of unmatched persons reduce. ODD matching should thus take about half the number of calculations as for stochastic matching.

Although many regions greatly the numbers of calculations required for stochastic and ODD matching, the smaller pools may raise concerns about the numbers of potentially mismatched couples at the end of each pool. For most practical purposes, probability-weighted and best of n matching are likely to be much quicker.

6.23 Conclusions

With the test data used here, immediate probability-weighted matching with sample sizes of about 30 gave good fidelity. The sample size can be increased to increase fidelity, or reduced to reduce runtimes. Immediate probability-weighted matching appears to be robust and flexible.

Immediate best of n sampling, with sample sizes of between 4 and 6 depending on age, gave slightly worse fidelity than probability-weighted sampling with samples of 30, but potentially lower runtimes. Best of n sample sizes need to be carefully chosen, as over-large samples will give too many close matches.

Stochastic matching is inherently more complex than immediate probability-weighted matching. In this case a large ad-hoc adjustment to the regression constants for males under 60 was needed to avoid improbable matches between old males and young females. With this adjustment, stochastic matching gave slightly worse fidelity than probability-weighted sampling with samples of 30.

Batch matching by randomized age order gave lower fidelity than stochastic matching, and it is difficult to see how it can be generalized to take into account variables other than age.

ODD matching gave too many moderate and high probability marriages, with almost no low probability marriages. A modified version of ODD using probability weighting performs much better, but it is still inferior to stochastic matching.

Immediate matching methods seem likely to give similar or better fidelity than batch matching, be valid with small and large regions, and give lower runtimes.

7. THE CUMPSTON MICROSIMULATION MODEL OF AUSTRALIAN HOUSEHOLDS

The Cumpston model was constructed for this thesis, to test proposed microsimulation techniques, examine data sources available in Australia, test model calibration and validation techniques, and assess the feasibility of more capable models.

The model begins with a 1% sample from the 2001 Australian census. After omitting visitors and incomplete households, about 175,000 persons remain, a 0.9% sample.

Data arrays are maintained for dwellings and individuals, with cross-references allowing all the individuals in a dwelling or family to be identified. The array structure allows ready addition of new dwelling and person data fields.

Education, employment, occupation, earnings and retirement savings are simulated for each person. These help in the simulation of household formation and housing. Disease incidence and development are simulated for each person.

A wide range of validation measures, largely based on those used in the Orcutt, CORSIM, DYNACAN and APPSIM models, are employed.

From runtimes for the Cumpston model, an estimate is made of the runtime for a similar model with 2.2m initial persons, allowing for some savings from multithreading.

7.1 Objectives of the Cumpston model

The objectives of the model are

- Testing the simulation techniques proposed in this thesis, particularly sampling with loaded probabilities, event alignment by random selection, and best of n matching
- Looking at the data sources available for microsimulation modelling in Australia
- Testing error-detection, calibration and validation techniques
- Assessing the feasibility of more capable models, with many more persons and geographic subdivisions.

The model is not intended to be comprehensive - for example, it does not cover social security, expenditure and investments. To save programming and validation effort, regression models have generally been simplified by using sex and age as explanatory variables, rather than using separate models for different sex and age groups.

7.2 Input data

Initial data came from the HSF (Household Sample File), a 1% sample of persons and households from the Australian 2001 census. This consisted of three flat files, for households, families and persons, with common household identification numbers allowing linking. The HSF was based on the place of residence on census night, so the following omissions were made to obtain data approximately based on usual place of residence

- Visitors from within Australia
- Visitors from outside Australia
- Persons in hotels and motels.

Table 7.1 shows the numbers of each type of omission.

Table 7.1 Input data for the Cumpston model

Source	Persons
HSF	188,013
less visitors from within Australia	-5,849
less visitors from outside Australia	-1,877
less persons in hotels/motels	-1,479
less families with only one partner	-3,449
less lone parents with no children	-315
Input data	175,044

Families with one partner, or with a lone parent and no children, were clearly incomplete, and were omitted. For confidentiality reasons, the HSF omits any household members beyond the first 6, so about 1,000 persons may have been lost for this reason (ABS 2003a p1). The estimated resident population of Australia at 30/6/01 was about 19.5m, so the 175,044 persons in the input data were a sample of about 0.9% of the population.

7.3 File structures and cross-references

Early attempts for this thesis to use databases with Visual Basic and C# had resulted in runtimes at least 20 times longer than those obtained with array storage. It was thus decided to use arrays to store person and dwelling data, together with the cross-referencing needed for a range of purposes. This meant that much of the cross-referencing provided automatically by databases had to be manually coded and tested. Once debugged, however, this cross-referencing proved stable and versatile. Additional dwelling and person data fields can readily be added, as shown by the addition of weighted records in 2009, and care and disease modelling in 2010.

7.4 Dwelling data

Table 7.2 shows the fields recorded for each dwelling. Storage bits for each field are from Albahari & Albahari (2008 p21).

Table 7.2 Dwelling data

Field	Type	Bits	Remarks
Dwelling number	int	32	index
Date occupancy	float	32	
First person in dwelling	int	32	cross-reference
Mortgage	int	32	
Mortgage repayment per month	int	32	
Number of families in dwelling	short	16	
Number of persons in dwelling	short	16	
Place in region pool	int	32	cross-reference
Region	int	32	
Rent per week	int	32	
Status	Boolean	1	
Structure	short	16	
Tenure	short	16	
Type	short	16	
Value	int	32	
Weight	int	32	
Total excluding index		369	

Each of the data items for dwellings is stored as a one-dimensional array, indexed by the dwelling number. The person number of the first person in the dwelling is used as a cross-reference to the data for all the persons in the dwelling. A similar structure appears to be used in INHASIM (Inagaki 2009a p4).

7.5 Dwelling pools

Three separate dwelling pools are maintained for each of 57 statistical divisions, the second highest subdivision level used by ABS for census purposes in 2001. These pools are for occupied private dwellings, vacant private dwellings, and for non-private dwellings. Each pool is a one-dimensional array containing dwelling numbers, together with a scalar variable giving the number of dwellings in the pool. Any newly constructed private dwelling is added to the end of the relevant array. When a vacant dwelling becomes occupied, is it added to the end of the occupied array, and the dwelling at the end of the vacant array is moved into the empty place in that array. Each array thus has no gaps. The place of each dwelling in its pool is recorded as one of the dwelling fields.

The main purpose of these pools is to allow quick searches for suitable dwellings. For example, when a family is modelled as moving between statistical divisions, the pool of vacant private dwellings in their destination is used to search for suitable dwellings.

7.6 Person data

Table 7.3 shows the fields recorded for each dwelling.

Table 7.3 Person data

Field	Type	Bits	Remarks
Person number	int	32	index
Age group	short	16	
Annuity flag	Boolean	1	
Annuity pa	int	32	
Area of dwelling	short	16	
Business effect	float	32	
Business flag	Boolean	1	
Business years	short	16	
Care level	short	16	
Date birth	float	32	
Disease code for 5 conditions	short	80	
Disease date for 5 conditions	float	160	
Disease number of conditions	short	16	
Disease severity	float	32	
Dwelling	int	32	cross-reference
Education year	short	16	
Employee flag	Boolean	1	
Family	short	16	
Highest schooling level attained	short	16	
Immigration data	float	32	
Immigration pending	Boolean	1	
Income code	short	16	
Income	int	32	
Income from business	int	32	
Income from other sources	int	32	
Income from wages	int	32	
Industry code	short	16	
Next person in dwelling	int	32	cross-reference
Occupation code	short	16	
Place in person pool	int	32	cross-reference
Previous person in dwelling	int	32	cross-reference
Qualification code	short	16	
Rank in immigrant family	short	16	
Sex	short	16	
Status	Boolean	1	
Study code	Boolean	16	
Study old code	short	16	
Superannuation flag	Boolean	1	
Superannuation fund	int	32	
Superannuation pension flag	Boolean	1	
Superannuation smsf flag	Boolean	1	
Type	short	16	
Wage effect	float	32	
Wage years	short	16	
Total excluding index		1016	

Access to all the persons in a dwelling is obtained using the first person (stored for the dwelling) and the next persons (stored for each person). For example, the total household income can be calculated by starting at the first person, and moving to each

next person in turn, adding incomes. The dwelling number is recorded for each person. The previous person in each dwelling is recorded to allow consistency checks.

7.7 Person types

Table 7.4 shows the different person types used.

Table 7.4 Person types in input data

Type	Number
Partner	77,726
Lone parent	7,427
Child	54,499
Related person in family	7,797
Unrelated person in family	1,416
Lone person	16,471
Group member	5,912
Non-private resident	3,796
Total	175,044

ABS conventions have been followed in defining person types and family members. For example, the birth of a child to a child in the family results in the new mother becoming a lone parent in a new family in the household. A lone parent must have a resident child. Lone persons must be the only occupant of their dwelling, and group members can only live with other group members.

Family relationships are inferred, rather than explicitly recorded. A child living in a family which includes two partners is assumed to be the child of those partners. Similarly, a child living with a lone parent is assumed to be a child of that person. This creates difficulties when modelling the return of adults to live with their parents, as the identity of their parents is not recorded. Consideration is being given to adding explicit parent-child links, which would also allow better treatment of mixed families, and of inheritances. A linking method similar to Inagaki's (2011 p3-5) would be used.

7.8 Person pools

Separate person pools are maintained for each combination of 8 areas, 8 person types and 9 age groups (576 pools in all). Like the dwelling pools, each person pool is a one-dimensional array, together with a scalar giving the number of persons in the pool. The person number of any person changing pools is added to the end of the new array, while its place in the old array is filled by the person number at the end of that array. The place of each person in its pool is recorded as one of the person fields.

The person pools are used for

- Sampling with loaded probabilities
- Event and person alignment
- Searching for suitable household members (for example, a male may search for a partner amongst adults resident in a particular area, who are not already partners).

To allow for age changes, the person pools are reconstructed at the start of each projection year.

7.9 Types of event simulated

Table 7.5 shows the main types of event simulated, and the units affected. Although immigrants are simulated as households, they may join existing households in Australia. Exits from households are simulated using “exit leaders”, who may be accompanied by “exit followers”. Exits may establish new households, or join other existing households. Partnership formations involve the exit of a person from a dwelling to partner with a person who is living another dwelling.

Table 7.5 Types of event simulated

Event	Unit	Chapter
Birth	Person	8
Death	Person	8, 14
Immigration	Household	9
Emigration	Person	9
Exits	Person	9
Moves between private dwellings	Household	9
Moves between private and non-private	Person	9
Education & training	Person	10
Employment	Person	11
Occupation	Person	11
Wages & business incomes	Person	11
Superannuation contributions & assets	Person	12
Housing	Household	13
Housing construction & demolition	Dwelling	13
Disease incidence and development	Person	14

7.10 Aligning to pool totals

Morrison (2006 p6-7) notes that a 90-year DYNACAN run involves more than 22,000 alignment pools, and more than 53 million prospective events. Assuming these are annual pools, there are about 250 pools a year, with about 2,400 prospective events per pool. The initial size of the DYNACAN sample is over 200,000 persons.

In the Cumpston model, alignment was carried out separately for 8 areas (each state and the Australian Capital Territory). Births in each cycle were aligned for 4 age groups (15-24, 24-34,35-44,45-54), and deaths, emigration and immigration were aligned over 9 age groups (0-14,15-24, …75-84,85+). The percentages of persons of each type at the end of each cycle were aligned over 8 person types (partner, lone parent, child, other related person, unrelated person, lone person, group member & non-private resident), 8 areas and 9 age groups. Table 7.6 shows the numbers of pools.

Table 7.6 Alignment pools

Description	Number areas	Number ages	Number types	Number pools
Births	8	4	1	32
Deaths	8	9	1	72
Emigrants	8	9	1	72
Immigrants	8	9	7	504
Person types	8	9	8	576
Total				1256

Numbers of persons, and births, deaths, emigrants and immigrants were aligned against national projections (ABS 2003c). As the necessary details by age group and area were not published, they were obtained from a deterministic projection program very closely replicating the national population projections. These approximate alignment totals led to some additional emigrants and immigrants, and moves between areas, in order to achieve exact alignment.

7.11 Use of Cumpston model for testing simulation techniques

It became clear that some of the commonly used simulation techniques were deficient, or inadequate for some potential applications. Theoretical analyses, and Monte Carlo simulations using hypothetical data, were sufficient to identify some problems and suggest improvements (see chapters 3, 4 and 6). But full testing of suggested microsimulation techniques generally requires extensive simulations with a realistic model of many households. The model's first practical use was thus as a testbed for proposed techniques, particularly sampling with loaded probabilities, event alignment by random selection, and best of n matching. The model also proved useful for testing microsimulation with weighted records, as proposed by Dekkers.

7.12 Unexplained heterogeneity

Wolfson (2009) notes that

> "…in contemporary microanalytic multivariate statistical estimation, it is rare that more than 10% of the observed variation or heterogeneity in real populations is 'explained' by the independent variables … this unexplained heterogeneity can still be incorporated explicitly into the modeling by characterizing its distribution."

This observation is true for most of the linear and logistic regressions in this thesis. For many of the linear regressions, the standard errors for the predicted values were found to be similar for each data point, all lying close to the root mean square error for the regression. In such cases, the goodness of fit was tested by generating simulated values including randomly distributed normal values with zero mean and standard deviation equal to the root mean square error. The percentile distribution of these simulated values was then graphically compared with the actual values. Figure 12.1 is an example of this technique.

The observed root mean square errors were also used to impute initial data values. For example, initial superannuation fund balances were imputed for each person with a fund using the regression model in table 12.1, adding randomly distributed normal values with zero mean and standard deviation equal to the root mean square error, then exponentiating.

7.13 Testing the validity of logistic regressions

Extensive use of logistic regression was made in finding statistical relationships, both for creating baseline data and projecting forward. An issue was the use of significance tests to decide whether particular coefficients should be retained in the model, and whether the overall fit was reasonable.

Hosmer and Lemeshow (2000 p16) note that the Wald test can behave in an aberrant manner, often failing to reject the null hypothesis when a coefficient is significant. They recommend the use of the likelihood ratio test, which assumes that twice the

improvement in the log-likelihood resulting from the addition of n variables is chi-square distributed with n degrees of freedom.

Logistic regression fits parameters by maximising the sum of the log-likelihoods, and this sum is routinely printed by the STATA software used here. For each logistic regression described here, the log-likelihood reduction is given, calculated as the percentage reduction in the log-likelihood from the fitted model, compared with a model fitted with just a constant.

7.14 Use of logistic regression rather than multinomial logistic regression

Many of the events modelled have multiple possible outcomes - changes of person type, changes of occupation, changes of region of residence. Initial trials with multinomial regression gave results that were hard to interpret, and computationally expensive to implement. Hosmer and Lemeshow (2000 p277-278) quote Begg and Gray's 1984 result that sequentially fitting binary regressions gives estimated regression coefficients that are consistent with those from multinomial regression, and under many circumstances the loss of efficiency is not too great. Hosmer and Lemeshow comment

> *"It has been our experience that the coefficients obtained from separately fit models are close to those from the multinomial fit"*

Based on the initial trials and this advice, logistic regression was used throughout. In some cases the probabilities of no change are high, so that computation times were reduced by modelling this probability first, and only if necessary simulating other transitions by calculating the probability of each separately and sampling from their sum.

7.15 Use of graphs for data exploration, testing and communication

Heavy use has been made of graphs in exploring data and testing fitted models. These graphs have shortened the thesis, and should make it easier for readers to understand the data and results. In general, very simple graphs have been used.

Plotting all the actual and fitted data points can create large documents, and hard to understand graphs. For this reason, frequent use has been of percentile or decile graphs, figure 12.1 being an example of a percentile graph. Figure 6.7 is another example, showing the ratio of simulated to observed marriages in each probability percentile. In this form of graph, a perfect fit gives small random variations around a flat line with value 1.0.

Grouping logistic regression results into percentiles or deciles by ranking on estimated probabilities is recommended by Hosmer & Lemeshow (2000 p147-151). This advice has been followed, using graphs rather than tabulations.

7.16 Validation of household microsimulation models

Caldwell and Morrison (1998 p1) commented

> *" …use of microsimulation models for policy analysis raises several relatively unexplored issues, among which one looms especially large. To what extent*

are microsimulation-based estimates credible as inputs to policy development? This issue is generally treated under the general term 'model validation'."

As Morrison (2008 p11-12) noted

> *"Longitudinal microsimulation models are complex pieces of software. Even in the best of circumstances … such models are susceptible to errors".*
>
> *"…The appropriate response to the prospect of errors in the model is a proactive approach for identifying and correcting errors before they emerge in the model's impact analyses".*
>
> *"…Model validation is necessarily an ongoing process."*

7.17 Validation of the Cumpston model

The Cumpston model was primarily developed to test simulation techniques, and to assess the feasibility of much larger models. Nevertheless, many of the validation techniques used in the Orcutt, CORSIM, DYNACAN and APPSIM models have proved relevant. A model intended for multiple purposes, many of them as yet unknown, needs to be reasonably close to reality in each of its processes. In summary, the main validation techniques used were

- A widely used programming language (C#) with strong type and name controls (Percival 2007)
- Program procedures intended to carry out common functions, such as sampling from cumulative probability distributions (Morrison 2008 p22)
- Stepping through code to check the accuracy of several sample cases for each mathematical procedure
- All assumptions contained in a single input file, with row and column headings corresponding with program variable names and indices (Morrison 2008 p19)
- Adjustments to baseline data to approximately correct non-responses and biases, and provide finer codes and quantitative values (Morrison 2008 p13-15)
- Use and retention of STATA Do files for extraction of baseline data and survey results
- Traps to detect loops, particularly in search processes
- Options to list details and event probabilities for each person simulated to have an event (Orcutt et al 1961 p303)
- Options to list details and matching probabilities for each person or dwelling considered as a potential match
- Options to cease running and dump output files of dwellings and persons, as soon as cases with error conditions are detected
- Options to provide output files of dwellings and persons at the end of the run, showing any error conditions
- Adjustment of input assumptions to ensure that projected time series and cross-sectional values are reasonably consistent with actual or expected experience, sometimes using iterative processes
- Optional use of event and state alignment to ensure that selected projected values equal actual or expected experience
- Output of time series and cross-sectional data into a single Excel file providing a comprehensive set of comparative graphs (Caldwell & Morrison 1998 p5, Morrison 2008 p32, Harding, Keegan & Kelly 2010 p52)
- Tests that yearly changes in superannuation accounts balance against cash flows and investment earnings
- A statistical test to detect creep between multiple runs

- Very close replication of national population projections (ABS 2003c)

7.18 Use of C#

An initial version of the Cumpston model was written in Visual Basic, using Excel files for input and output. 50-year projection results from this model, using a 0.1% sample of Australian households, were described by Cumpston (2007). Following Percival's 2007 recommendation of C# for APPSIM, the model was rewritten in C#, with input and output to CSV files. Percival's recommendation was based on the object-oriented nature of C#, and on its design to assist programmers develop code that is safer and easier to maintain. He noted that its runtime performance was not as fast as some alternatives, including C++.

C# proved to be an easy language to use, with its name prompt facility allowing the ready use of informative variable names. Its strict type and name control was helpful in maintaining accuracy. The ability to use dummy arrays as procedural arguments was useful in creating general purpose statistical routines. Jagged arrays were helpful in dealing with areas with large and small populations.

7.19 Use of STATA and CSV files

The STATA versions of the HSF files were used, and STATA was extensively used to derive statistical patterns, and to prepare CSV files of dwellings and persons for input to the Cumpston model. STATA Do files were kept for all operations, intended to allow rapid updating when future surveys or census samples became available. STATA readily outputs CSV files, and such files have no practical size limits, unlike Excel files. The STATA Do files for HILDA data proved easy to extend as updated versions of the HILDA survey became available each year.

All data output from the Cumpston model is similarly to CSV files, avoiding size limits. Where necessary, such files can be readily input to Excel or STATA for analysis. The main output files from the current version are

- A listing of all the dwellings in the data, optionally output at the end of the run
- A listing of all the persons in the data, optionally output at the end of the run
- A listing showing each of the occupied and vacant private dwellings in each region, and each occupied non-private dwelling, optionally output at the end of the run
- Optional event listings.

7.20 Event listings and error traps

For each type of event simulated, an output file was created giving sufficient details to verify the probability calculations and trace the consequences of each event. These files were large, and an option to suppress their creation gave significant time savings. A similar option was included in the model described by Orcutt et al (1961 p303).

One difficulty with Monte Carlo modelling of households is that the sequence of events leading up to a program failure can be hard to reconstruct. For this reason, many error traps have been inserted into the code. Whenever a potential error condition is detected, the program can output the current dwelling, person and event files, and cease execution. As an alternative, potential errors can be listed in dwelling and person files output at the end of the run.

7.21 Use of comparative graphs for validation

Key time series and cross-sectional results are output into a single CSV file, which is then pasted into a single Excel file providing a comprehensive set of comparative graphs. As many different changes can affect validation results, it is important that comparative graphs be readily updated. A similar file of validation graphs is used by APPSIM. The Excel file also contains consistency checks, for example that yearly changes in superannuation accounts balance against cash flows and investment earnings.

7.22 Statistical test for creep between multiple runs

Caldwell & Morrison (1998 p26-27) mention the use of dispersion measures obtained from 300 plus runs of a DYNACAN small sample, and from seven CORSIM runs. In adjusting input assumptions to ensure that projected time series and cross-sectional values from the Cumpston model were reasonably consistent with actual or expected experience, averages of 10 runs were typically used. For some results, creep was noticed, with the values stabilizing after a few runs to values noticeably different to the values from the first run. Investigation usually showed that creep was due to failure to initialize variables at the start of each run.

To help detect creep, a statistical test is calculated for each key result for each projection year as

((result for last run) - (result for first run)) / (2 * standard deviation)

Values of this test with magnitudes greater than unity are treated as worth investigation. The creep tests are built into the Excel file providing comparative graphs.

7.23 Runtimes for Cumpston model

Table 7.7 shows 50-year runtimes, without and with disease simulations. The figures are the results of single runs, and there are random variations between the times for the same processes without and with diseases. No time for deaths is shown for the "with diseases" case, as deaths are simulated as part of the "disease development" process.

Table 7.7 50-year runtimes in seconds

Process	Without diseases	With diseases
assumption input	0.34	0.34
births	7.29	7.41
creation of dwellings	1.15	1.15
creation of pools	1.90	1.81
data input	2.90	3.03
deaths	0.45	0.00
disease creation	0.00	5.71
disease development	0.00	7.46
disease incidence	0.00	6.75
dwelling updates	0.66	0.72
education	4.68	4.49
emigrants	0.53	0.53
exits	4.79	4.80
immigrant generation	0.81	0.86
immigrants	0.72	0.72
moves	3.96	3.98
non-private/private moves	0.33	0.34
private/non-private moves	0.31	0.25
summaries	0.14	0.11
superannuation	10.20	10.26
work	24.51	24.23
other	0.35	0.39
Total	66.03	85.35

Times for each main process are measured by programmed time recorders at the start and end of each process. Locations for these time recorders have to be carefully chosen, as each runtime use of a recorder involves a significant time cost. Recorders are thus placed at the start and end of blocks of code affecting all records (for example generation of all births for a year) rather than in code affecting individual records. Once this is done, overall runtimes are found to be little affected by the use of process time recording.

7.24 Potential runtime savings from multithreading

Fredriksen, Knudsen & Stolen (2011 p5-7) report that multithreading reduced the runtime of MOSART by a factor of 4-5 when 6 threads were employed, with far less effect from further threads. These results were obtained with a Linux-based server using 16 processors and 256 GB RAM. They note that steps involved in household formation are cumbersome to multithread, due to often subtle interactions between individuals, with little or no effect on runtime.

Of the processes in table 7.7, only disease creation, disease development, disease incidence, dwelling updates, education, superannuation and work seem likely to give significant gains from multithreading. Together these account for 70% of the total runtime for the model with diseases, so that a 75% reduction in these processes from multithreading might give an overall runtime reduction of about 52%.

Note that gains of this magnitude are unlikely to be obtained with the multithread processors typically found in personal computers. The overall heat output from such processors is monitored, and the cycle time slowed down to maintain safe operating temperatures.

7.25 Planned extension to 2,214 areas

A planned extension is to the 2,214 SA2 statistical areas introduced in the 2011 census, rather than the present 57 statistical divisions. Techniques for simulating increasing population in existing areas, and the spread into new areas close to cities, have already been developed for deterministic projections (Cumpston 2002). The household/dwelling matching methods developed in this thesis allow appropriate dwellings to be identified for moving households, taking into account their incomes and assets.

To ensure that there enough persons in each of the 2,214 areas to give reasonably representative results, about 1,000 persons per area may be needed, giving about 2.2 million persons in all. These persons will probably be synthesized from 2011 census cross-tabulations for the 2,214 statistical areas, using the 1% census sample to ensure realistic family structures.

Allowing for the 1016 bits per person estimated in table 7.3, for 8 bits per byte, and for a three-fold expansion of the number of person records during simulation, about 800 Mbytes of storage are likely to be needed for person records. Even allowing for dwelling records, output totals and more stored details per person, this seems well within the capacity of personal computers.

Allowing for the increase in initial population from 0.175m to 2,2m, and for a 50% runtime reduction from multi-threading, the 50-year runtime for a model including diseases might be about 9 minutes.

7.26 Conclusions

Sampling with loaded probabilities, alignment by random sampling and best of n matching all work well, and should give runtimes linearly proportional to the numbers of person records.

The Cumpston model also showed that microsimulation with weighted records works, but the runtime savings may not justify the additional program complexity.

A wide range of ongoing validation measures are needed to ensure that microsimulation estimates are credible for their intended purposes. A single well-annotated input file for assumptions, and a single output file allowing generation of extensive comparative graphs, proved particularly helpful.

A 50-year test run with disease simulation takes about 85 seconds. It is amongst the fastest of the models for which run times have been obtained. Its speed has been helped by sampling with loaded probabilities and immediate best of n matching.

A planned extension of the Cumpston model is to 2,214 statistical areas. This may require initial data of about 2.2 million persons, synthesized from census cross-tabulations. Assuming multithreading is used where feasible, and some increase in functionality, a 50-year projection may take about 10 minutes.

8. FERTILITY AND MORTALITY

Australian total fertility rates rose following the introduction of a baby bonus on 1/7/04, and have fallen slightly since the bonus was means-tested on 1/7/08. The bonus reversed the long-term decline in total fertility rates.

Age-specific fertility rates have fallen slightly below age 30, and risen above that age. Multiple births rose as a result of assisted reproduction technology, but fell sharply with the introduction of single-embryo transfer. Male to female birth ratios have been reasonably stable at about 1.054.

A logistic regression model for fertility was fitted to the 2001 census sample data, based on the 2,053 children born in the year prior to the census. Age, partnership status, numbers of prior children, the gap since the last child and education were all significant explanatory variables.

Adjustments to the fitted fertility model were made to give approximately the right numbers of children in each maternal 5-year age group in the first projection year, and approximately the right numbers of children in each projection year, after allowing for the known changes in age-specific fertility rates.

Examination of other national models suggests that multiple equations are generally needed to do justice to the many factors affecting fertility.

Numbers of deaths were projected using 00-01 mortality rates, adjusted in line with the changes observed in mortality rates for 5-year age groups for each sex from 00-01 to 08-09.

Australian data show that married persons have much lower mortality rates than other persons. UK data show that persons in skilled occupations have much lower mortality rates than those in unskilled occupations. No use has yet been made of marital or occupation differentials in the Cumpston model.

An alternative set of death projections was made using mortality rates for each of 584 disease stages. Disease-based mortality assumptions do not appear to have been used in other national models.

8.1 Structure of Cumpston model

Births are simulated using sampling with loaded probabilities amongst females aged 15-54. Multiple births, and the sex of each child, are simulated. If a non-partnered female is simulated to have a child, her type is changed to “lone parent”, and she becomes the head of a new family.

Deaths are simulated in two different ways. If diseases are not being simulated, deaths are simulated assuming sex and age related mortality rates incorporating the progressive reductions evident for most age groups, using loaded probabilities. If diseases are being simulated, each stage of each disease has its own assumed mortality rate, there are no “normal” deaths, and all-case sampling is used.

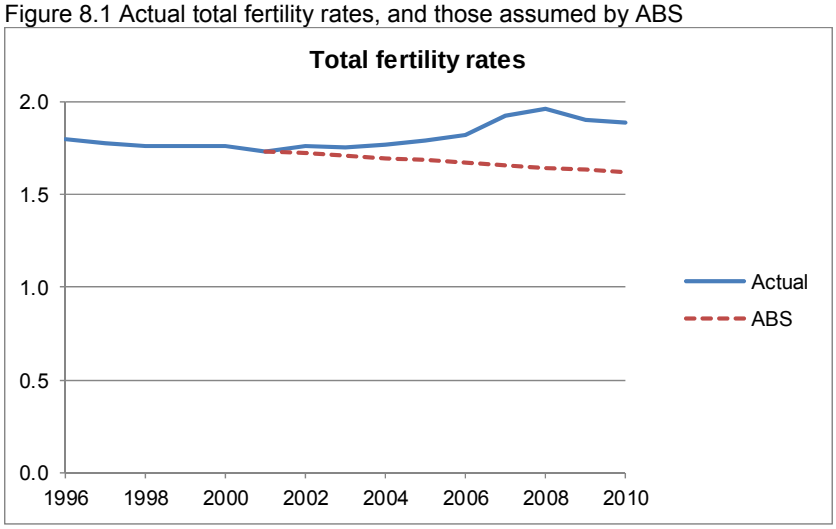
Although there are strong mortality differentials between partnered and unpartnered persons, and between social classes, no attempt has yet been made to allow for these differentials in simulations.

8.2 Australian total fertility rates

ABS calculates age-specific fertility rates as the number of births during a year to women of an age, multiplied by 1000 and divided by the number of women of that age in the estimated resident population at 30 June of that year. Total fertility rates are calculated as the sum of the age-specific rates, divided by 1000 (ABS 2011c p57).

Figure 8.1 shows the actual total fertility rate from 1996 to 2010 (ABS 2011c table 1.9) compared with the medium rates assumed in national population projections published in 2003 (ABS 2003c). A $3,000 cash bonus for the birth of child was introduced on 1/7/04. The then Treasurer, Peter Costello, quipped that a third child could be "for the country" (Jansen & Dill 2009). The baby bonus rose to $5,000 on 1/7/08, and was then indexed and mean-tested. Higher fertility rates may also have been partly due to increased immigrant numbers.

Figure 8.1 Actual total fertility rates, and those assumed by ABS



8.3 Age-specific fertility rates

Table 8.1 shows that fertility rates from ages 15 to 29 dropped slightly, while those from age 30 on rose (ABS 2011c table 1.1). Although there were large rises in the fertility rates from age 40 to 49, these ages only contributed about 4% of total births in 2010.

Table 8.1 Age-specific fertility rates

Year	15	20	25	30	35	40	45	TFR
2000	17.7	59.2	107.9	109.5	48.7	8.7	0.4	1.76
2001	17.7	58.0	104.4	107.9	49.0	9.2	0.4	1.73
2002	17.2	56.4	104.5	111.2	52.1	9.7	0.4	1.76
2003	16.1	54.2	102.5	112.3	54.2	10.0	0.5	1.75
2004	16.0	52.7	101.8	114.0	57.2	10.5	0.5	1.76
2005	15.7	51.8	102.0	117.0	60.3	10.8	0.5	1.79
2006	15.3	51.4	101.0	120.4	63.4	11.3	0.6	1.82
2007	16.0	55.5	105.6	125.9	67.8	12.6	0.7	1.92
2008	17.2	56.5	104.9	127.1	70.6	14.1	0.7	1.96
2009	16.7	54.0	102.4	124.0	68.7	14.2	0.7	1.90
2010	15.5	52.5	100.2	123.3	69.7	14.8	0.7	1.89

In projections with the Cumpston model, assumed changes in age-specific fertility rates after 01-02 were based on the actual changes for each 5-year age group shown in table 8.1.

8.4 Multiple births

Figure 8.2 shows multiple birth percentages from ABS (2011c, table 9.9). Laws & Hilder (2008, p22) attribute the overall increasing trend in multiple births in the last two decades to

> *"the increased use of fertility drugs and assisted reproduction technology, delay in childbearing and the growing number of older mothers"*

Jansen & Dill (2009) note that the fast-increasing practice of single-embryo transfer has resulted in a sharp fall in multiple deliveries after IVF.

Figure 8.2 Multiple births as a % of total confinements

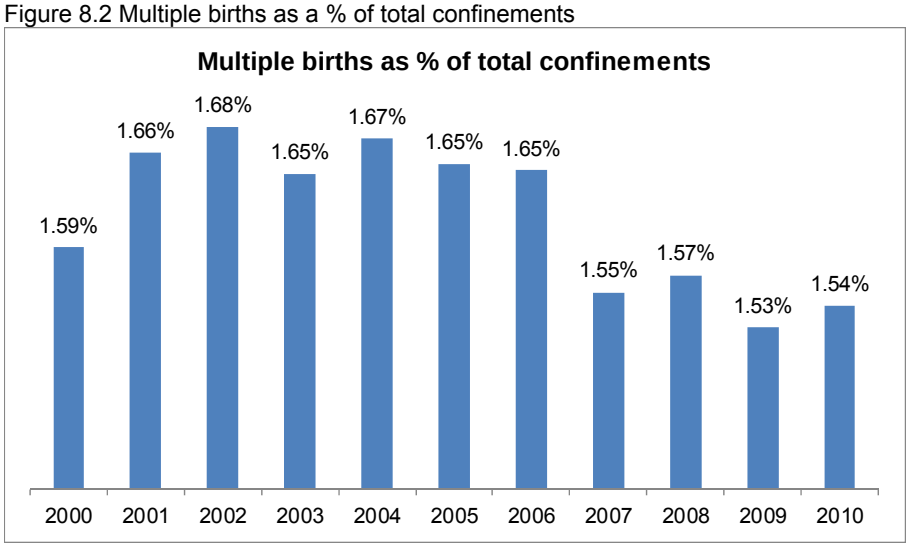


Table 8.2 shows the multiple birth rates assumed in the Cumpston model. The rates to 31/12/06 are based on the observed averages from calendar years 2001 to 2006, and those from 1/1/07 on are based on the averages for 2007 to 2010.

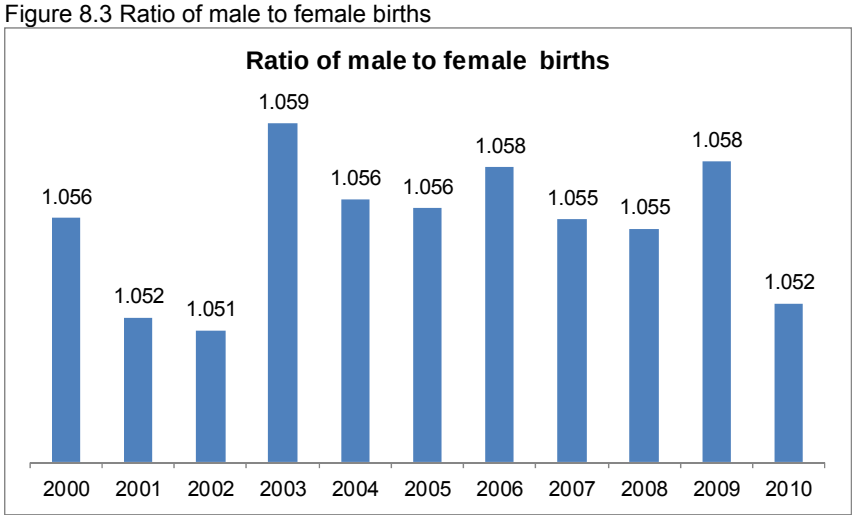
Table 8.2 Fertility rates as a percentage of those in 2001

Number children	To 31/12/06	From 1/1/07
1	0.98339	0.98453
2	0.01627	0.01522
3	0.00030	0.00022
4	0.00004	0.00003
Total	1.00000	1.00000

8.5 Sex ratio for births

The sex ratios in figure 8.3 are from the Australian Bureau of Statistics (2011c, table 1.9). The average ratio of 1.054 over the 10 years to 2010 was assumed.

Figure 8.3 Ratio of male to female births



8.6 Fitted fertility model

The logistic model in table 8.3 was fitted to data from the 2001 census sample file. There were 49,071 females aged 15 to 49, recorded as partners or lone parents living in the first family in a household. Of these females, 2,053 had children born in the year prior to the census.

Table 8.3 Logistic regression model for fertility in 2001

Variable	Coeffi-cient	Standard error	Probability level
age1718	1.620	0.626	0.010
age19	2.554	0.607	0.000
age20	2.131	0.585	0.000
age25	2.221	0.584	0.000
age30	1.930	0.585	0.001
age35	1.049	0.587	0.074
age40	-0.442	0.596	0.458
age45	-1.690	0.628	0.007
partner	1.631	0.080	0.000
child1	2.735	0.323	0.000
child23	1.679	0.323	0.000
gap1	-2.125	0.333	0.000
gap234	-0.840	0.323	0.009
gap5	-1.653	0.328	0.000
qalhigh	0.123	0.055	0.024
constant	-6.157	0.578	0.000
N			49071
LLR			20.1%

The explanatory variables used in the adopted model were

- age1718 for females 17 or 18, age 19 for females 19, age20 for females 20-24, then in 5-year bands to 45-49
- partner, 1 if in partnership, 0 otherwise
- child1, 1 if one previous child living in the family
- child23, 1 if 2 or 3 previous children living in the family

- gap1, 1 if the youngest child living in the family was born in the 2nd year before the census
- gap234, 1 if the youngest child living in the family was born 3rd, 4th or 5th year before the census
- gap 5, 1 if the youngest child living in the family was born in the 6th year before the census, or earlier
- qalhigh, 1 if female's non-school qualification diploma or higher.

The loglikelihood reduction for the adopted model was 20.1%. Note that this model is for the probability of giving birth in a year, and multiple births have to be separately modelled.

Being in a partnership increases the probability of giving birth by about 5.1 times (calculated as the exponential of the regression coefficient). Having one previous child greatly increases the probability of birth, but the effect is smaller for multiple previous children. Children are unlikely to be born in the year after a year in which a birth has occurred, or more than 5 years after. Having a diploma or higher qualification may increase the probability of birth by about 13%.

Kippen, Gray & Evans (2005) used 2001 census data to show that mothers with 2 children of the same sex are more likely to have a third child than mothers with a boy and a girl. No evidence of this effect was found here, but this may reflect the use of a 1% sample rather than the whole census.

No allowance was here made for children no longer living with their mother, or for children in the household not born from the mother. These should not be major problems, provided the projection method is similarly based on children living in the family, regardless of parentage.

No attempt was made to include the mother's income or work status in the fitted model, as these might have been changed by the birth. These variables could have been examined using the HILDA longitudinal data, as was done by Pennec & Bacon (2007 p33-38).

8.7 Adjustments to the fitted fertility model

Table 8.4 shows age adjustments to the fitted fertility model, derived by comparing actual numbers born in the year ending 30/6/02 with the averages of 10 simulation runs for that year.

Table 8.4 Age adjustments to fitted fertility model

Age	15-	20-	25-	30-	35-	40-	45-
Adjustment	1.394	1.425	1.217	1.377	1.383	1.121	0.185

Table 8.5 shows projection year adjustments to the fitted fertility model, derived by comparing actual numbers born in each projection year with the averages of 10 simulation runs, after making the adjustments in table 8.4.

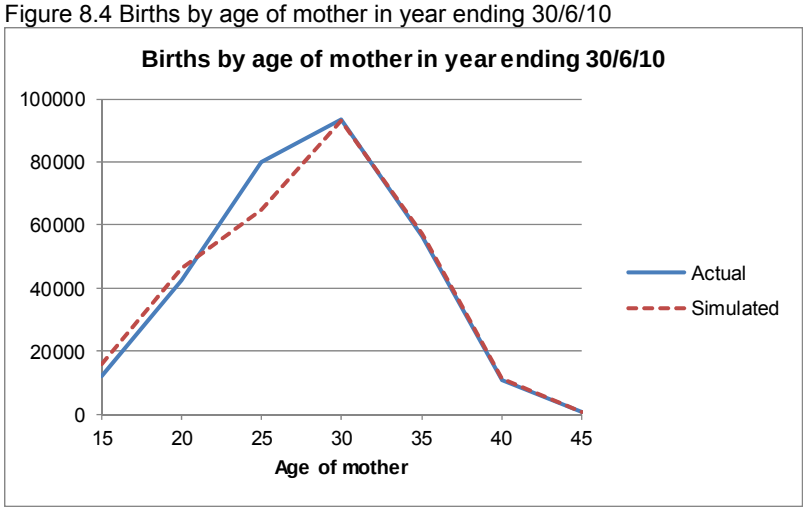
Table 8.5 Projection year adjustments to the fitted fertility model

Year	1	2	3	4	5	6	7	8	9
Adjustment	0.986	0.961	0.946	0.953	0.953	0.978	0.981	0.975	0.970

8.8 Comparisons between actual and projected births in 09-10

Figure 8.4 shows that simulated birth numbers in the year ending 30/6/10 were reasonably close to actual, except that simulated births to mothers ages 25-29 are 19% below actual.

Figure 8.4 Births by age of mother in year ending 30/6/10

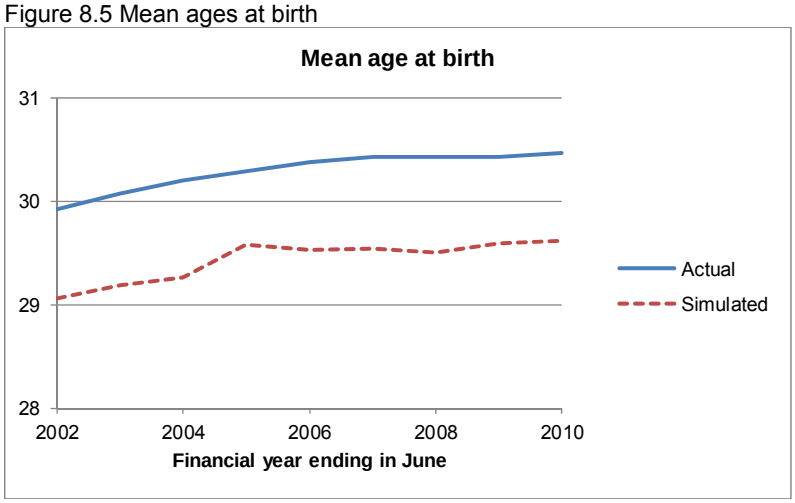


The overestimation of births for females aged 15-24, and the underestimation of births to mothers aged 25-29, may largely reflect failure to correctly model partnership status for young females.

8.9 Mean ages at birth

Figure 8.5 shows that the simulations are underestimating the mean age at birth by just under a year. This appears to be largely due to the overestimation of births for females aged 15-24, and the underestimation of births to mothers aged 25-29. Note that actual mean ages are uncertain, as they have been estimated from data on the numbers of births for each maternal 5-year age group.

Figure 8.5 Mean ages at birth



8.10 Fertility processes in other national models

Table 8.6 briefly describes the fertility processes used in five other national models. In general, multiple equations seem needed to do justice to the many factors underlying fertility.

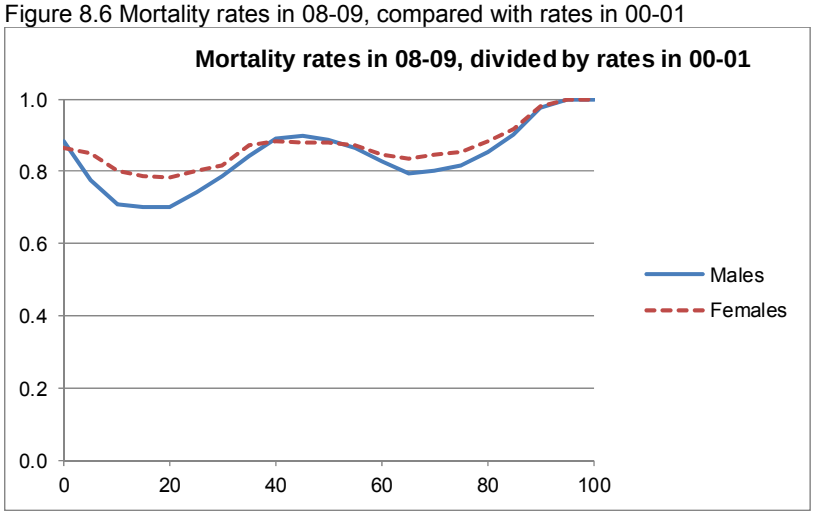
Table 8.6 Other national models

Model	Description	Reference
APPSIM	Separate regression models for partnered and unpartnered females without children and with one child, and for females with 2 and 3 children. Covariates include age, age squared, migrant status, married, partnered, education level, labour force status, years since last birth and age mother at first birth	Pennec & Bacon 2007 p33-38
DYNASIM3	Seven-equation parity progression model; varies based on marital status; predictors include age, race, education, work history, marriage duration, number of prior births, time since last birth; uses vital rates after 39	Favreault & Smith 2004 p5, p7
INHASIM	Marital fertility rates by parity and mother's age, with no allowance for births outside marriage	Inagaki 2009a p5
SESIM III	Separate regression models for females 18-49 who have moved from their parents and have and have not given birth to a child; covariates age, marital status, pensionable income quartile, indicator for market work, highest education, age of youngest child	Flood et al 2005 p30
SVERIGE	Separate regression models for married & unmarried females, with explanatory variables age, age squared, 2 education variables, 5 income variables, one at work variable and a partnership variable for unmarried females	Holm et al 2002 p17-18

8.11 Mortality rates in 08-09, compared with rates in 00-01

Figure 8.6 shows estimated mortality rates in 08-09, divided by estimated rates in 00-01, for 5-year age-groups for males and females. Mortality rates were estimated from the numbers of deaths in each calendar year (ABS 2010c table 1.9), divided by the estimated resident population at June 30 (ABS 2010b table 9).

Figure 8.6 Mortality rates in 08-09, compared with rates in 00-01

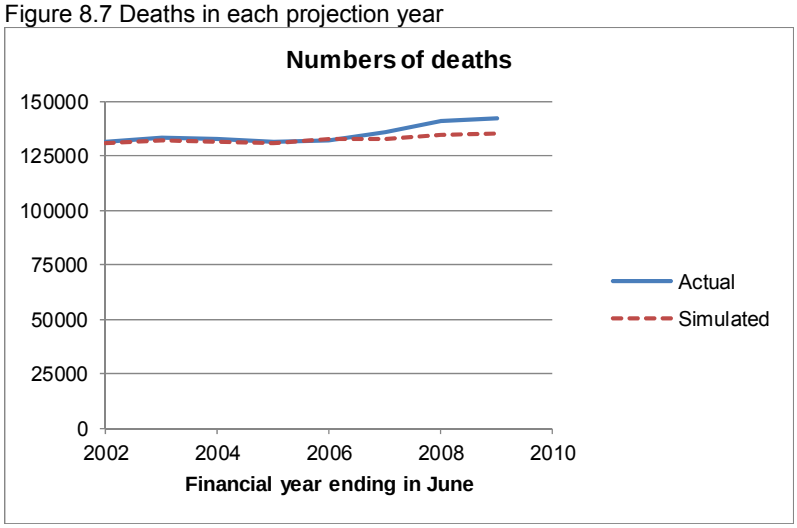


The estimated ratios were smoothed, by averaging the estimated rate with the rates for the two five-year age groups below it, and the two age-groups above it. For ages 95-99 and 100+, actual rates in 08-09 appeared to be higher than in 00-01, but the erratic nature of the exposure data suggested the estimated rates were unreliable. Actual rates for ages 95-99 and 100+ were thus assumed to be identical in 00-01 and 08-09.

8.12 Actual and simulated deaths in each projection year

Figure 8.7 shows actual deaths in each financial year to 08-09, together with those simulated from the actual mortality rates in 00-01 and the assumed mortality rate improvements in figure 8.6. Simulated values are close to actual for all but the last 3 years. The reasons for this divergence are not clear, but may be related to uncertainties in the exposure and death data at old ages, and to age biases in the 2001 census sample.

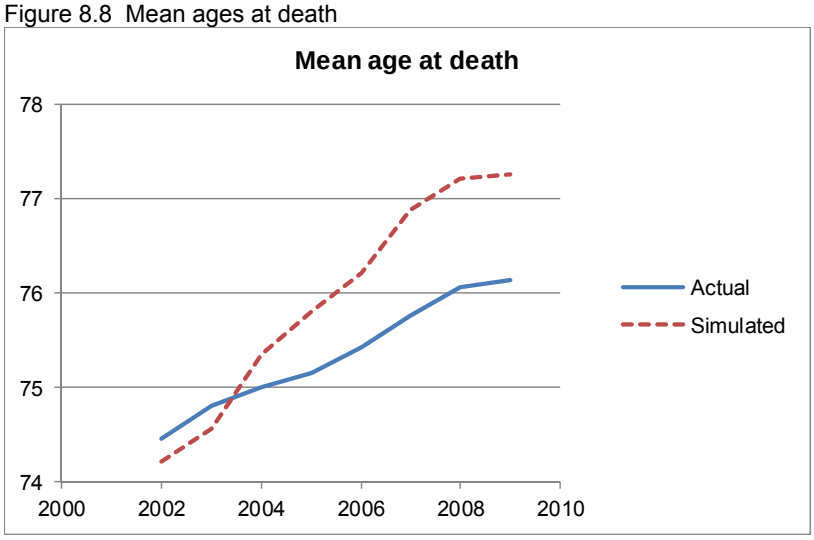
Figure 8.7 Deaths in each projection year



8.13 Mean ages at death

Figure 8.8 shows that the simulated mean age at death is close to actual for the first two projection years, but then grows more quickly than the actual mean, being about 1.1 years higher for the last 3 projection years. Again, the reasons for this divergence are not clear. Note that actual mean ages are uncertain, as they have been estimated from data on the numbers of deaths for each 5-year age group.

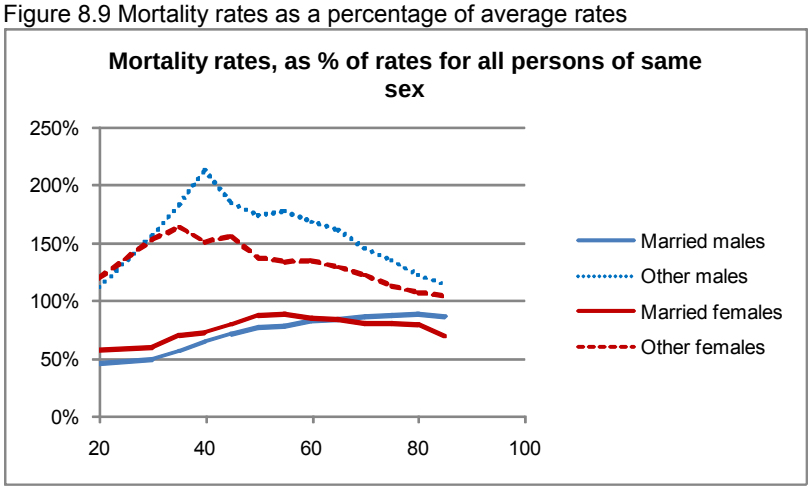
Figure 8.8 Mean ages at death



8.14 Possible mortality rate adjustments for marital status

The ratios in figure 8.9 were obtained from ABS death statistics (2003f), including information on the marital status of persons age 20 and over dying in 2001. No use of mortality differentials by marital status has yet been made in the Cumpston model.

Figure 8.9 Mortality rates as a percentage of average rates



The results are in general agreement with the conclusion by Gardner & Oswald (2004) that

> *"For men, the effect of being married is positive and substantial. For example, it approximately offsets the large negative consequences of smoking. For women, the influence of marriage is smaller, at half the size of the smoking effect".*

8.15 Potential mortality rate adjustments for social class

The life expectancies in table 8.7 are from the UK National Statistics Office (2007). The life expectancies for all males and all females are similar to those for Australia in 2001, and life expectancies for each social class in Australia were assumed to be the Australian all person life expectancies, plus the UK differentials. Mortality loadings were fitted by minimising the squared differences between the actual and fitted life expectancies at ages 0 and 65, for each sex and social class.

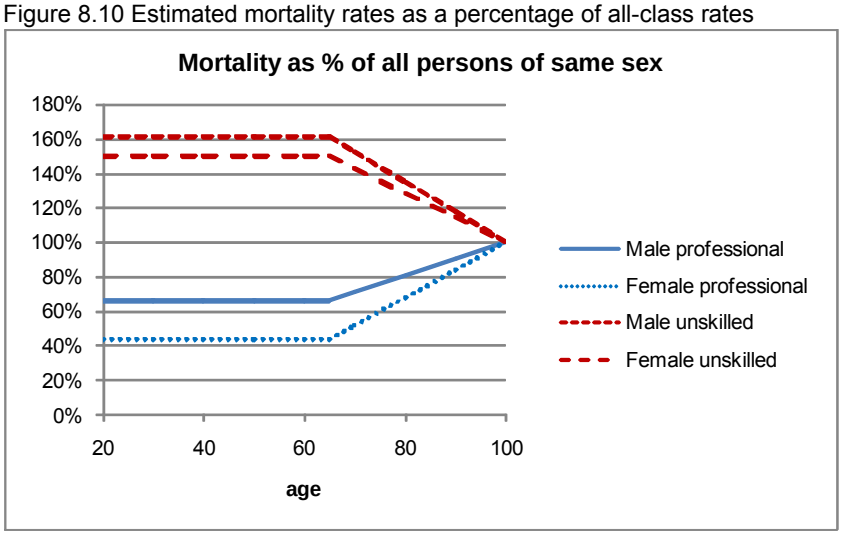
Table 8.7 UK life expectancies 2002-05, and fitted Australian mortality loadings

Sex & social class	At birth	At 65	Mortality loading
Males			
Professional	80.0	18.3	-34%
Managerial & technical/intermediate	79.4	18.0	-28%
Skilled non-manual	78.4	17.4	-16%
Skilled manual	76.5	16.3	6%
Partly skilled	75.7	15.7	18%
Unskilled	72.7	14.1	61%
Unclassified	73.8	15.1	41%
All males	77.0	16.6	0%
Females			
Professional	85.1	22.0	-57%
Managerial & technical/intermediate	83.2	21.0	-33%
Skilled non-manual	82.4	19.9	-17%
Skilled manual	80.5	18.7	12%
Partly skilled	79.9	18.9	17%
Unskilled	78.1	17.7	50%
Unclassified	77.9	17.6	53%
All females	81.1	19.4	0%

Figure 8.10 gives estimated mortality rates, as percentages of the all-class rates, for Australian professional and unskilled persons, separately for males and females. Social class loadings were assumed to zero before age 20, to be constant from 20 to 65, and then to decrease linearly to zero at age 100. This linear decline in mortality differentials after age 65 has been found by the author to apply in two specialized occupational groups - an order of nuns, and the Fellows of the Australian Academy of Technological Sciences and Engineering.

Relationships between mortality and social class in Australia have recently been examined by Clarke & Leigh (2011). They found the relative risk of mortality between the poorest and richest income quintile was 1.88 times higher.

Figure 8.10 Estimated mortality rates as a percentage of all-class rates

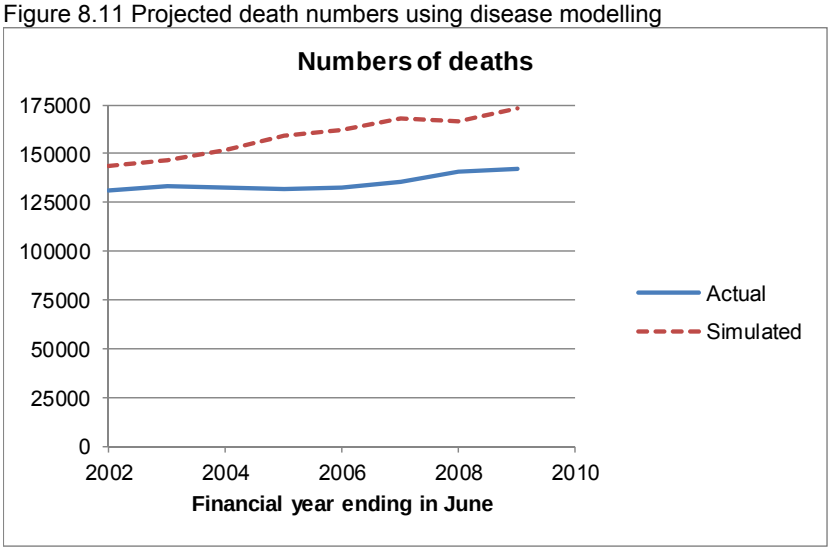


No use of mortality differentials by class has yet been made in the Cumpston model.

8.16 Death numbers using disease modelling

Figure 8.11 shows simulated death numbers using the assumed mortality rates for each disease stage (see chapter 14). The simulated death numbers are 9% higher than actual in 01-02, rising to 22% higher in 08-09.

Figure 8.11 Projected death numbers using disease modelling



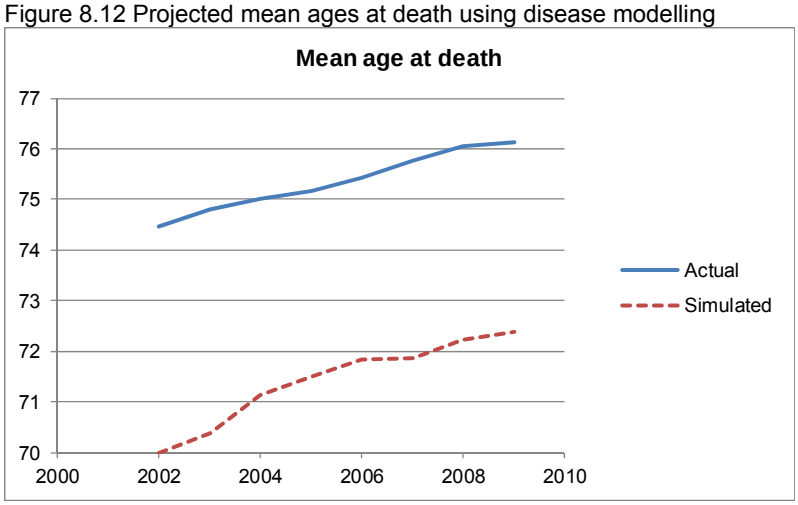
The assumed disease incidence and mortality rates are largely based on data before 2000, so lower death rates may apply to some common diseases, and to deaths from external causes. For some diseases, mortality rates are expressed as ratios of normal population mortality, so that allowances for improving mortality are automatically made in the projections. Other diseases have fixed mortality assumptions, and these assumptions may need to be reduced as treatment or general health improves.

To help make long-term projections of mortality from different disease types, trend lines can be fitted to long-term past data on death causes, and reasonable assumptions mad about future trends. This was done by Cumpston (2001).

8.17 Mean ages at death using disease modelling

Figure 8.12 shows simulated mean ages at death using the assumed mortality rates for each disease stage. The simulated ages at are about 4 years less than actual ages.

Figure 8.12 Projected mean ages at death using disease modelling



The low simulated mean ages suggest that mortality rates from some diseases are being underestimated at older ages. Expert advice is needed on likely future mortality rates from diseases.

8.18 Other national models

Table 8.8 gives five examples of the treatment of mortality in other models.

Table 8.8 Other national models

Model	Description	Reference
APPSIM	Probability by age & sex, as assumed in ABS population projections. Inclusion of further differentials is under investigation.	Pennec & Bacon 2007 p23
DYNASIM3	Three equations; time trend from Vital Statistics 1982-97; includes socio-economic differentials; separate process for the disabled based on age, sex & disability duration	Favreault & Smith 2004 p5
INHASIM	Probability by age & sex, allowing for mortality improvements	Inagaki 2009a p5
SESIM	Probability by age & sex, allowing for mortality improvements; allowing for income & marital status ages 30-64, and for marital status & highest level of education from 65 on	Flood et al 2005 p30
SVERIGE	Probability by age & sex, allowing for mortality improvements; allowing for education, income, work status & marital status at ages 25-29, for education & marital status at 60 to 74, and for marital status at ages 75+	Holm et al 2002 p16

8.19 Conclusions

A logistic fertility model was fitted to 2001 census data, adjusted to replicate age-specific fertility rates in the first projection year, and to replicate total numbers of children in each projection year. In spite of these adjustments, the model overestimated births to females aged 15-24 in 09-10, underestimated births to females aged 25-29, and underestimated mean ages at birth by just under a year. These differences may largely reflect failure to correctly model partnership status for young females.

A mortality model based on 00-01 mortality rates, and on subsequent changes in rates for 5-year age groups, gave simulated values falling below actual in the last 3 projection years. Mean ages at death were close to actual for the first two projection years, but then grew more quickly, being about 1.1 years higher in the last 3 projection years. These differences may be related to uncertainties in the exposure and death data at old ages, and to age biases in the 2001 census sample.

An alternative mortality model, based on assumed mortality rates for 584 disease stages, projected death numbers about 9% higher than actual in the first year, rising to 22% higher in 08-09. Simulated mean ages at death were about 4 years lower for each projection year. These differences may reflect the use of pre-2000 disease data, and the need for expert advice on likely future mortality rates from diseases.

9. MIGRATION, MOVES AND HOUSEHOLD EXITS

The Cumpston model simulates immigration into Australia, emigration out of Australia, moves within and between 57 statistical divisions, and transitions between 8 different person types.

A single event can lead to multiple changes in person types and residences. For example, a male partner may leave a household, taking with him one of the partnership's children, and move to a residence interstate. Both partners would change type, to become lone parents.

The available census and longitudinal data are not adequate to estimate reliably the many different probabilities involved in immigration, emigration, migration and household exits. Large adjustments to estimated probabilities were needed to approximately replicate observed household types.

This chapter shows that sampling with loaded probabilities (chapter 3) can be used to model emigration, moves and household exits, and that immediate best of n matching (section 6.7) can be used to match partners, parents and children, other related persons, unrelated persons and group members.

9.1 Structure of Cumpston model

Immigrants and emigrants are assumed to have the age composition assumed in ABS projections (2003c), with numbers adjusted to balance with subsequent experience. Immigrants are assumed to have the family composition and personal characteristics of persons recorded in the 2001 census as arriving in the previous year. Emigrants are assumed to emigrate as individuals rather than families.

Moves are defined as instances where an entire family, a lone person or a group moves to another dwelling. An exit is where one or more persons leave together, but at least one person remains behind. Probabilities of movements to other statistical divisions are simulated, based on the age of the move or exit leader, the distance to the destination and the number of persons at the destination.

Some types of household exits involve joining a household at the destination - for example, a person living as a child in one household, moving to become a partner in another household. Similarly, some types of immigrant families may join an existing household, rather than create a new one.

For each person aged 15+, the model simulates exit leadership, and change of household type if an exit occurs. The person types simulated are partner, lone parent, child, other related person, unrelated person, lone person, group and non-private resident. For all types of exit leader except lone person and non-private resident, the model simulates persons of equal or lower type sometimes following the leader.

Destination households are selected using immediate "best of n" matching. For example, 10 unpartnered persons of the opposite sex in the destination region are randomly selected as potential partners for an exit leader seeking a partner, and the one with the best regression score chosen.

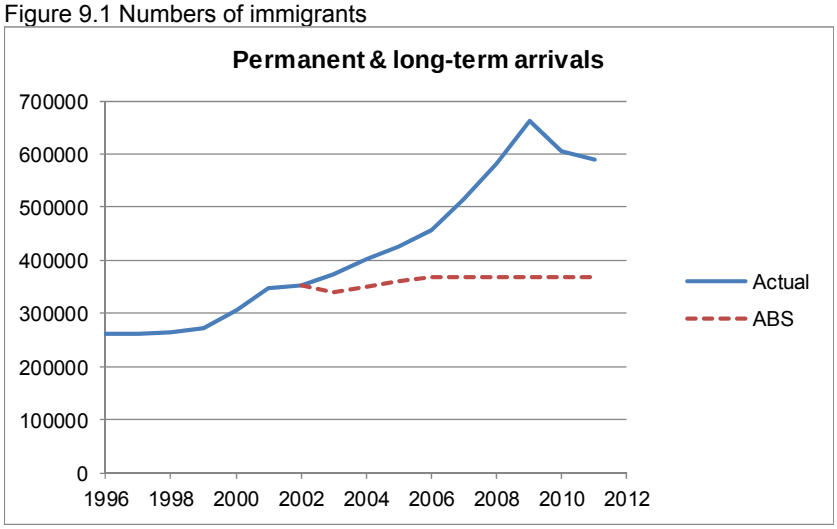
Persons are simulated to exit from private to non-private dwellings as individuals. Persons exiting non-private residences are assumed to have the type distribution of persons of their sex and age group.

Emigrants, moves, exit leaders, and moves between private and non-private dwellings are simulated using loaded probabilities.

9.2 Immigration

Figure 9.1 shows actual numbers of permanent and long-term arrivals in Australia, where a long-term arrival is someone planning to stay more than 12 months (ABS 2011d table 1). Also shown are medium immigration assumptions in national population projections (ABS 2003c p23). Labour needs from the economic boom for most of the decade to 2010, and much higher levels of overseas students coming to Australia, resulted in numbers of immigrants greatly exceeding ABS's 2003 assumptions.

Figure 9.1 Numbers of immigrants



In the Cumpston model, immigrants are closely based on the composition by age, sex and destination state assumed in the ABS projections but their total numbers are based on actual numbers. Based on the type composition of persons recorded in the 2001 census sample as arriving in Australia in the year before the census, their types are randomly allocated. For each person, their date of birth, highest level of schooling, post-school qualification, industry, occupation and income are randomly allocated, again based on recent immigrant data.

All persons are assumed to form new private households, or to join existing private households. At the start of each projection year, the generated immigrants are grouped into households, matching persons in the following order

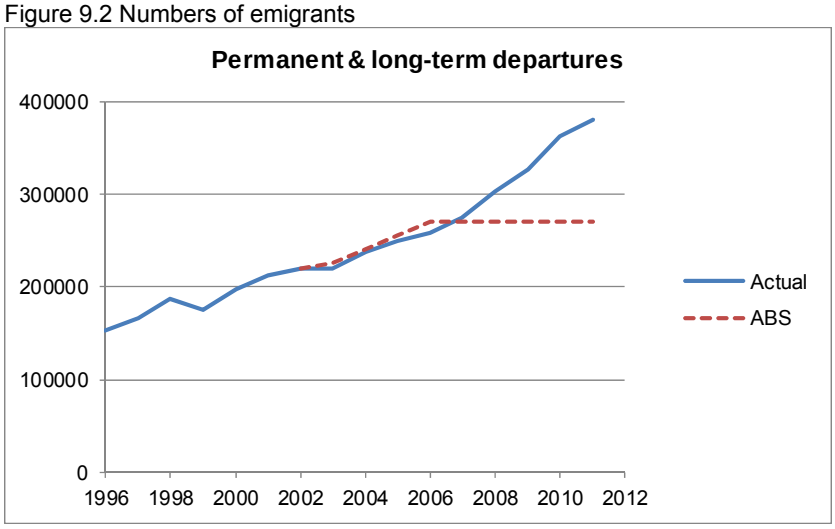
- partners
- lone parents
- children
- other related persons
- unrelated persons
- group members
- lone persons (the remaining persons).

As immigrants are simulated for an area during each cycle in the projection year, immigrant households are randomly selected from the generated households for the area, removing them from the immigration queue for the year. All the generated immigrant households are used during the year, spread as evenly as possible across the cycles.

9.3 Emigration

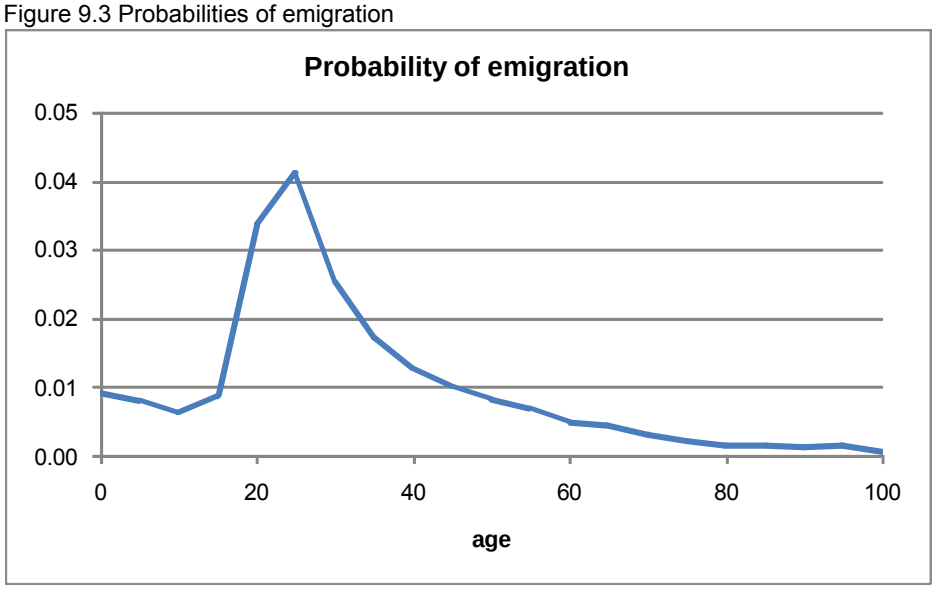
Figure 9.2 shows actual numbers of permanent and long-term departures from Australia (ABS 2011d table 1). The continuing growth in emigrants is likely to have resulted from the growth in immigrants, as some permanent immigrants and many overseas students leave after a few years.

Figure 9.2 Numbers of emigrants



The Cumpston model assumes the probabilities of emigration in figure 9.3, which are based on the medium emigration assumptions in the national population projections (ABS 2003c p23). These probabilities are assumed to apply to all person types other than lone parents, and persons are assumed to emigrate as individuals rather than families. To avoid problems with children remaining behind, lone parents were assumed not to emigrate. Data on the family composition of emigrants are needed to help make more realistic assumptions.

Figure 9.3 Probabilities of emigration



9.4 The HILDA longitudinal survey

The HILDA Survey is funded by the Department of Families, Community Services and Indigenous Affairs, and run by the Melbourne Institute of Applied Economics and

Social Research (Watson 2008). It is a longitudinal survey, attempting to follow the 19,914 persons included in the 7,682 households interviewed for wave 1 of the survey, as well as other persons who subsequently share households with the original sample. The fieldwork for wave 1 began in August 2001, and was largely completed by the end of 2001. Similarly, the fieldwork for waves 2 to 6 was largely completed by the end of 2002 to 2006.

Each person in the HILDA has a unique identifier, which remains constant throughout the survey. These cross-wave identifiers were used to merge the start and end HILDA files for each of the five years, identifying 9,764 losses from the survey, and 7,307 entering or re-entering. 5,396 of the losses were of entire households, presumably as a result of survey non-response. Table 9.1 shows the results.

Table 9.1 Non-response and exit rates for HILDA

Type of person	Wave1-2	Wave2-3	Wave3-4	Wave4-5	Wave5-6	Total
At start wave	19914	18295	17691	17209	17469	90578
Losses	2750	2167	1941	1468	1438	9764
Continuing	17164	16128	15750	15741	16031	80814
New	1131	1563	1459	1728	1426	7307
At end wave	18295	17691	17209	17469	17457	88121
Non-responses	1828	1304	1100	572	592	5396
Non-response	9.2%	7.1%	6.2%	3.3%	3.4%	6.0%
Exits	922	863	841	896	846	4368
Exit rate	4.6%	4.7%	4.8%	5.2%	4.8%	4.8%
Exits still in survey	522	343	275	315	297	1752

Exits were defined as persons leaving households, where at least one other person remains in their household, and the household responds to the next interview. When this occurs, at least one of the identifiers of the persons present at the start of the wave appears at the end of the wave, and any missing persons from the household can be inferred to be exits rather than non-responses. Of the 4,368 exits, 1,752 were still in the survey at the end of the year, but in other households.

9.5 Movements of households

A household movement is here defined as the movement of all the persons in a dwelling to another dwelling. Table 9.2 shows the assumed probabilities of movement of the four types of person considered able to lead the movement of a family from one dwelling to another. Children, other related persons and unrelated persons are considered to follow movement leaders, rather than initiate moves.

Table 9.2 Assumed probabilities of movement leadership

Age	15-	25-	35-	45-	55-	65-	75-	85-	Moves
Partner	0.38	0.21	0.12	0.07	0.07	0.05	0.03	*0.03*	2245
Lone parent	0.41	0.27	0.15	0.10	0.06	0.03	0.02	*0.02*	451
Lone person	0.48	0.31	0.15	0.11	0.08	0.05	0.04	0.04	1182
Group member	0.25	0.21	0.13	0.12	0.06	*0.03*	*0.02*	*0.02*	44
Total									3922

These probabilities were derived from the first 6 waves of the HILDA longitudinal survey. For partners, attention was restricted to families where the partners had stayed in the one dwelling, or had moved to another dwelling together. For group members, attention was similarly limited to cases where at least two of the members

had stayed in the one dwelling, or had moved together. In all, 3922 persons were identified as movement leaders.

The assumed movement probabilities are those derived from the HILDA data, with the values in italics smoothed slightly. Persons moving were identified as those with a non-zero distance moved. Note that movement leadership is simulated separately for each partner, and the other partner follows their move, so that the probabilities of partners moving are about double those in table 9.2. Similarly, movement leadership is simulated for each member of a group, so that the probability of a two-person group moving is about double those in table 9.2.

9.6 Distances travelled by moving households

The average distances travelled in table 9.3 are from the same HILDA data. Cases where fewer than 10 persons moved have been omitted from the table. Lone persons may be less willing to move long distances as they become older. The 44 moving group members in the data are not enough to reach any reliable conclusions about the distances they are likely to move.

Table 9.3 Average kilometres travelled by moving households

Age	15-	25-	35-	45-	55-	65-	75-	85-	Total
Partner	113	227	210	202	146	214	354		198
Lone parent	77	63	119	154	28				97
Lone person	288	255	181	201	168	74	61	111	218
Group member	148	74	26						91

9.7 Proportions of movement leaders changing major statistical region

HILDA has 13 major statistical regions (MSRs). NSW, Victoria, Queensland, SA and WA are split into capital city and rest of state regions. Tasmania, NT and the ACT are each a single region. The proportions in table 9.4 were obtained from the 3,922 persons classed as movement leaders. There is no clear evidence of any age pattern in table 9.4, so that the total percentages were assumed to apply. For example, a lone parent who moved was assumed to have a 10% chance of changing statistical divisions. As the seven largest MSRs are identical to the 7 largest statistical divisions, it was thought reasonable to apply the MSR total percentages to changes of statistical division. Better data would be helpful.

Table 9.4 Proportion of movement leaders changing MSR

Age	15-	25-	35-	45-	55-	65-	75-	85-	Total
Partner	15%	19%	18%	23%	24%	21%	25%	14%	19%
Lone parent	8%	10%	12%	12%	7%				10%
Lone person	18%	16%	13%	16%	17%	17%	19%	13%	16%
Group member	10%	6%	30%						12%

9.8 Model for selecting destination statistical division

Table 9.5 gives the results of logistic regression models fitted to the 20.208m persons recorded at the 1996 or 2001 censuses, where their usual statistical division of residence was known at both at the census and 5 years previously. These data were supplied by ABS, subdivided into 10-year age groups. Distances between each pair of

statistical divisions were measured as great circle distances between their population centroids, as described by Cumpston Sarjeant (1998).

Table 9.5 Logistic regression coefficients for selection of destination

Age group	const	ln(kms)	ln(size)	LLR
1	-6.396	-0.880	0.668	0.161
2	-8.044	-1.030	0.868	0.239
3	-8.099	-0.820	0.778	0.176
4	-7.762	-0.811	0.743	0.165
5	-6.589	-0.928	0.705	0.176
6	-3.011	-1.198	0.535	0.188
7	-2.506	-1.459	0.593	0.189
8	-4.456	-1.298	0.703	0.229
9	-4.899	-1.256	0.716	0.247
Total				0.187

A separate regression model was fitted for each 10-year age group, of the form

probability (mover selecting destination i) = exp(score) / (1 + exp(score))

where score = b_0 + b_1* ln(distance in kms) + b_2* ln(number persons in destination i)

Persons in the youngest age group would be likely to have the same characteristics as their parents, who would have been largely in age groups 3 and 4.

The regression coefficients for the logarithm of the distance are more negative with increasing age, suggesting that older persons are likely to travel to closer destinations when they change statistical divisions. Persons in age group 2 show a preference for larger destinations, while persons in age groups 6 and 7 appear to prefer smaller destinations.

Comparing the actual numbers of persons to each destination with the fitted numbers gave a range from 0.41 (Far West Victoria) to 5.05 (Darwin). Attempts were made to explain these variations using the additional variables labour force participation rate and climatic comfort, but these did not give noticeably better fits. The actual to expected ratio for each statistical division was thus used as a final adjustment to the derived probabilities, at all ages.

9.9 Probabilities of exit leadership

Of the 4,368 exits in table 9.1, 3,172 were identified as “exit leaders”, defined as the person of the lowest relationship type and highest age amongst those leaving a household. Table 9.6 shows the coefficients of the 6 regression models fitted for each type of person amongst these exit leaders. No single persons are included, as HILDA confidentiality procedures make it impossible to detect single persons moving to other households. Table 9.6 also shows the loglikelihood reductions for the models, and the numbers of leaders.

Table 9.6 Regression coefficients for probabilities of exit leadership

Regression parameters & LLR	Regression coefficient for relationship before exit					
	Partner	Lone parent	Child	Related person	Unrelated person	Group Member
sex	-0.508	-0.403				
multi	1.608	1.768				
age15			3.181			
age25	-0.831		3.181		0.580	
age35	-1.329		2.539	-0.830		
age45	-1.477		2.071	-0.830		-0.797
age55	-1.477	0.576	2.071	-0.830		-1.265
age65	-1.477	0.576	2.071	-0.830		-1.265
age75+	-0.659	0.576	2.071	-0.830		-1.265
constant	-1.818	-2.364	-4.998	-1.235	-1.273	-0.908
LLR	0.045	0.047	0.203	0.025	0.012	0.025
Leaders	1040	208	1287	238	68	331

Sex is recorded by HILDA as 1 for males and 2 for females. "Multi" is 0 for single families and groups, and 1 for multiple families. All coefficients were significant at the 5% level.

As an example of the use of these coefficients, the probability of a female partner aged 35-44 being an exit leader is calculated as

exp(-0.508*2 - 1.329 -1.818) / [1 + exp(-0.508*2 - 1.329 - 1.818)] ie 1.53%.

The negative coefficients for sex in the model from exit leadership from partnerships suggest that males leave the dwelling in the majority of cases where partnerships break up. Male lone parents appear to be more likely to leave their households than female lone parents. Sex appears to have little or no effect on probability of exit as a leader from other relationships.

The increases in exit rates from partnerships at age 75 plus are likely to reflect deaths or movements to institutions, rather than partnership separations.

Where no significant differences were found between age deciles, larger age groupings were used. For example, lone parents aged 55 plus were found to be more likely to leave their households than younger lone parents.

Partners and lone parents in multi-family households are about 5 times as likely to be exit leaders than similar persons in single-family households. This may reflect a general preference for single-family households, with multiple families a temporary expedient following migration or partnership formation. No significant difference was found between exit rates as leaders from same-sex and opposite-sex partnerships.

The HILDA data provide no information on the probabilities of lone persons joining another household. Lone persons were assumed to have half the exit rates of group members of the same age.

9.10 Adjustment factors for exit rate assumptions

The exit assumptions in table 9.6, together with the type distributions in table 9.8, were used to make projections to 30/6/10, and the resulting proportions by type of person

for each age group were compared with alignment proportions derived from ABS household projections (ABS 2004a). The adjustment factors in table 9.7 were chosen so as to reduce the differences between the projections and the alignment proportions.

Table 9.7 Adjustment factors for exit rate assumptions

Age	Partner	Lone parent	Child	Related person	Unrelated person	Lone person	Group member
15	2.00	1.00	0.10	1.00	1.00	2.00	0.50
25	1.00	2.00	1.00	1.00	1.00	2.00	0.50
35	1.20	2.00	1.00	1.00	1.00	1.00	1.00
45	1.00	2.00	1.00	1.00	1.00	1.00	1.00
55	0.10	0.10	1.00	1.00	1.00	1.00	1.00
65	0.10	0.00	1.00	1.00	1.00	2.00	1.00
75	0.00	0.00	1.00	0.10	1.00	2.00	1.00
85	0.00	0.00	1.00	0.10	1.00	2.00	1.00

Of the 56 adjustment factors, 33 are equal to 1.00, indicating that the derived exit rates seemed to be giving reasonable results. Given that the exit rates for lone persons were arbitrarily taken as half those for group members, it is not surprising that 5 of them had to be adjusted. Large reductions in exit rates had to be made for partners and lone parents aged 55+, children aged 15-24 and related persons aged 75+. It is not clear why these large reductions were needed, and the adjustment factors could usefully be reviewed when data become available from the 2011 census.

9.11 New type distributions for exit leaders

End types were only known for 1,348 of the 3,172 exit leaders in the HILDA data. Of the 1,348, 238 were partners with children living with them. The new type distributions in table 9.8 were obtained by tabulations from the 1,348, with some smoothing and adjustments to balance with alignment proportions.. Given that the end types were unknown for 58% of the exit leaders, there must be considerable doubt about the reliability of the distributions. Trials with multinomial logit models for type changes gave erratic results, perhaps reflecting the limited data.

Table 9.8 New type distributions for exit leaders

Type at start	Sex	Age group	Type at end 1	2	3	4	5	6	7
Partner parent	Male	15-	0.440	0.000	0.130	0.125	0.000	0.060	0.245
Partner parent	Male	25-	0.280	0.050	0.030	0.004	0.002	0.604	0.030
Partner parent	Male	35-	0.350	0.020	0.030	0.004	0.002	0.574	0.020
Partner parent	Male	45-	0.290	0.020	0.030	0.004	0.002	0.634	0.020
Partner parent	Male	55-	0.230	0.000	0.000	0.004	0.002	0.744	0.020
Partner parent	Male	65-	0.170	0.000	0.000	0.004	0.002	0.804	0.020
Partner parent	Male	75-	0.110	0.000	0.000	0.004	0.002	0.864	0.020
Partner parent	Male	85-	0.050	0.000	0.000	0.004	0.002	0.924	0.020
Partner parent	Female	15-	0.100	0.700	0.100	0.004	0.002	0.018	0.076
Partner parent	Female	25-	0.070	0.790	0.040	0.004	0.002	0.074	0.020
Partner parent	Female	35-	0.100	0.540	0.000	0.004	0.002	0.344	0.010
Partner parent	Female	45-	0.080	0.620	0.000	0.004	0.002	0.284	0.010
Partner parent	Female	55-	0.060	0.330	0.000	0.004	0.002	0.594	0.010
Partner parent	Female	65-	0.040	0.220	0.000	0.004	0.002	0.724	0.010
Partner parent	Female	75-	0.020	0.110	0.000	0.004	0.002	0.854	0.010
Partner parent	Female	85-	0.000	0.000	0.000	0.004	0.002	0.984	0.010
Other	Both	0-	0.000	0.000	1.000	0.000	0.000	0.000	0.000
Other	Both	15-	0.050	0.000	0.070	0.220	0.030	0.030	0.600
Other	Both	25-	0.430	0.000	0.020	0.020	0.010	0.450	0.070
Other	Both	35-	0.300	0.000	0.040	0.020	0.010	0.580	0.050
Other	Both	45-	0.200	0.000	0.040	0.020	0.010	0.730	0.000
Other	Both	55-	0.200	0.000	0.040	0.060	0.030	0.670	0.000
Other	Both	65-	0.200	0.000	0.000	0.060	0.030	0.710	0.000
Other	Both	75-	0.100	0.000	0.000	0.000	0.000	0.900	0.000
Other	Both	85-	0.100	0.000	0.000	0.000	0.000	0.900	0.000

There are some strong differences between male and female parents with children. For example, 29% of males aged 45-54 become partners in new households, and only about 2% become lone parents. By contrast, only about 6% of females become partners in new households, with 62% becoming lone parents.

No strong differences were observed between male and female exit leaders who were not partners with children, so common assumptions were made for them. To avoid programming difficulties, no persons who were not partner parents were assumed to become lone parents.

9.12 Logistic models for probability of following exit leader

The 3,172 exit leaders were accompanied by 1,021 “followers”. For each type of leader before exit, logistic models were fitted for the probability of following the leader. The regression coefficients for the adopted models, and the log-likelihood reductions obtained from the models, are shown in table 9.9. All coefficients were significant at the 5% level.

Table 9.9 Regression coefficients for probability of following leader

Regression parameters & LLR	Regression coefficient for leader type before exit					
	Partner	Lone parent	Child	Related person	Unrelated person	Group member
sexleader	0.218					
age15	-0.749	-0.861	-1.475			
age25	-0.749	-1.655	-1.475			-1.081
age35+	-1.786	-3.858		-2.423		-1.081
constant	-1.581	-0.708	-1.968	-2.587	-3.314	-2.018
LLR	0.076	0.212	0.057	0.060	0.000	0.035
Followers	660	190	81	25	4	61

For example, the probability of a child following an exiting female partner aged 25-34 is estimated as

$$\exp(0.218*2 - .749 - 1.581) / [1 + \exp(0.218*2 - .749 - 1.581)] \quad \text{ie } 13.1\%.$$

The probability of the child following an exiting male partner is similarly estimated as 10.8%. Both these probabilities seem low, and suggest that young children usually stay in the family dwelling on the separation of their parents.

9.13 Differences between simulated and ABS type/age proportions

Table 9.10 shows the differences between the proportions of persons of each type simulated for each age group, compared with those projected by ABS (2004a).

Table 9.10 ABS projections of type proportions for each age group at 30/6/10, less simulated

Type	0-	15-	25-	35-	45-	55-	65-	75-	85-
1	0.000	0.001	0.007	-0.009	0.010	0.039	0.007	-0.004	-0.007
2	0.000	0.008	-0.001	-0.022	-0.031	-0.022	0.011	0.024	0.038
3	0.019	-0.008	-0.005	0.017	0.010	0.004	0.000	0.000	0.000
4	-0.009	-0.004	0.002	0.001	-0.002	-0.008	-0.005	-0.012	0.004
5	0.000	-0.002	-0.014	-0.006	-0.001	-0.005	-0.006	-0.004	-0.002
6	-0.007	-0.001	0.010	0.014	0.008	-0.011	-0.008	0.003	-0.044
7	0.000	0.004	-0.003	0.002	0.009	0.008	0.006	0.004	0.005
8	-0.003	0.002	0.004	0.003	-0.001	-0.005	-0.005	-0.011	0.006

The differences seem reasonable for most type/age combinations, except for lone parents from age 35 on, and lone persons at 85+. The absolute error averaged across the 72 type/age combinations is 0.008.

9.14 Probability of exit leader changing statistical division

The probabilities of changing major statistical region in table 9.11 were derived from the 1,348 exit leaders in the HILDA data for which source and destination region were known. Leaders aged 15-44 seem less likely to change region than those younger or older.

Table 9.11 Probability of exit leader changing major statistical region

Age	Same MSR	New MSR	Probability
0-24	597	200	25%
25-54	445	65	13%
55+	31	10	24%
Total	1073	275	20%

The above probabilities for changes of major statistical region were assumed to apply to changes in statistical division for exit leaders. If they changed statistical division, the models in table 9.5 were used to randomly select the destination.

9.15 Regression model for partners derived from census sample data

The parameters in table 9.12 were estimated from a data file made up of 38,263 records of opposite sex partners from a 1% sample of the Australian 2001 census, plus the same number of unmatched pairs created by randomization. Note that the data are for all partnerships, not just recent ones.

Table 9.12 Regression coefficients for partnerships

Parameter	Estimate	SE
agediff^2	-0.0131	0.0001
agediff	0.0836	0.0017
hscdiff^2	-0.2253	0.0055
qalldiff^2	-0.0372	0.0010
constant	1.5688	0.0148
N		38263
LLR		0.372

A logistic model was used to estimate the probability of a male and a female, both known to be partners, of being partners with each other. The independent variables found to give useful reductions in the magnitude of the log likelihood were

- agediff, the age of the male less the age of the female
- agediff^2, the square of agediff
- hscdiff^2, the square of the difference in the codes for the highest level of schooling completed (these codes ranged from 1 for still at school to 5 for not stated)
- qaldiff^2, the square of the difference in the codes for non-school qualification (these codes ranged from 1 for postgraduate to 8 for not stated)

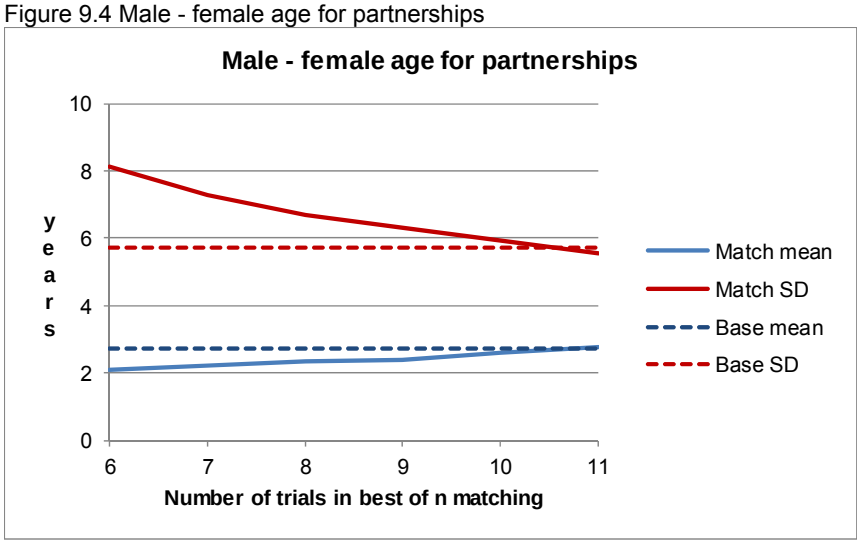
The agediff and agediff^2 parameter estimates suggest that the highest probabilities of being in partnership occur when the male is about 3.2 years older than the female.

9.16 Matching persons exiting households with new partners

As persons exit households, their new household type is simulated, and their destination statistical division is randomly simulated. Each person simulated to become a partner is then matched to a person in the destination, using immediate "best of n" matching from a sample of 10 unpartnered persons aged 15+ in the destination. If there are less than 10 unpartnered persons aged 15+, then a different destination is simulated. The sample size of 10 was based on the results in figure 9.4,

which shows that both the mean and standard deviation of the age difference for simulated partnerships with a sample size of 10 are very close to the observed values for partnerships in the 2001 baseline data.

Figure 9.4 Male - female age for partnerships



The exit leader joins the selected partner in their dwelling, together with any persons simulated to follow the exit leader. This may result in multiple-family households, for example when the selected partner is a child living with their parents. As the exit rates from multi-family households are high, this may only be temporary.

9.17 Logistic model of probability of two persons being parent and child

In the regressions in table 9.13, agediff is the age of the parent, less the age of the child. The parameter estimates suggest that the peak probability of a person being the child of a male occurs when the age difference is 32 years, and the peak probability for a female parent occurs when the age difference is 28.7 years.

Table 9.13 Regression coefficients for parent/child associations

Parameter	Male parent Estimate	Male parent SE	Female parent Estimate	Female parent SE
agediff^2	-0.0066	0.0001	-0.0091	0.0001
agediff	0.4227	0.0054	0.5232	0.0057
constant	-6.2343	0.0850	-6.8592	0.0799
N		44263		54224
LLR		0.099		0.130

The above parameters for male parents were estimated from a data file made up of 44,263 records of children with male parents from the 1% sample file from the 2001 census, and the same number of randomized records. Similarly, the parameters for female parents were obtained from 54,224 records of children with female parents from the sample file.

9.18 Logistic models of probabilities of other persons entering households

The results in tables 9.14 and 9.15 were obtained from data files made up of all the persons of the particular type, together with the same number of randomized records. Age difference was taken as the age of the reference person in the household, less the age of the person entering the household. The estimated coefficients for age difference in table 9.14 were adjusted so as to give matches with means and standard deviations of age differences similar to those observed in the 2001 census sample (see 9.19).

Table 9.14 Regression coefficients for other related and unrelated persons

Parameter	Other related Estimate	Other related SE	Other adjusted Estimate	Unrelated Persons Estimate	Unrelated persons SE	Unrelated adjusted Estimate
agediff			-0.0750	0.0070	0.0023	-0.0210
agediff^2	-0.0006	0.0000	-0.0006	-0.0005	0.0001	-0.0005
constant	0.3559	0.0296	0.3559	0.1559	0.0468	0.1559
N		7797			1416	
LLR		0.039			0.012	

The regression coefficients in table 9.15 gave good results for group matches, and no adjustments were made.

Table 9.15 Regression coefficients for group members

Parameter	Estimate	SE
agediff	0.0147	0.0024
agediff^2	-0.0021	0.0001
constant	0.3805	0.0292
N		5912
LLR		0.086

9.19 Selected numbers of trials for matching

Table 9.16 shows the numbers of trials selected for “best of n” matching when matching different types of person to households, and the resulting age differences. The age difference for partners is measured as the male age less the female age. For children and related persons, the age difference is measured between the person joining the family and the head of the family, where the head is defined as the oldest person of the highest type in the family (including the person joining the family). For unrelated persons and group members, the age difference is measured between the person joining the family and the head of the household, where the head is defined as the oldest person of the highest type in the household (including the person joining the household). When determining family or household heads, partners are treated as the highest type, with lone parents, children, other related, unrelated and group members following.

Table 9.16 Numbers of trials and age differences for different types of match

Type	Number trials	Match mean	Match SD	Base mean	Base SD
Partner	10	2.7	5.8	2.8	5.7
Parent/child	8	-31.2	6.5	-31.6	6.9
Other related	3	-9.3	19.2	-8.7	22.2
Unrelated	3	-8.9	17.4	-9.1	17.4
Group	6	-3.8	7.3	-3.8	7.1

For all 5 match types, means and standard deviations of age differences are close to the observed values in the baseline data.

9.20 Movements between private and non-private residences

The probabilities of movement from non-private to private households in table 9.17 were arbitrarily chosen, based on the decline in non-private residence for most age groups, and on the relatively low exit rates of persons in aged care. The probabilities of movement from private to non-private were then chosen so as to approximately replicate the observed proportions of persons in non-private residences shown in the 2001 census sample.

Table 9.17 Assumed probabilities of movement between private & non-private residences

Age	Private to non-private	Non-private to private
0-	0.000	0.40
5-	0.000	0.38
10-	0.008	0.36
15-	0.009	0.34
20-	0.004	0.32
25-	0.002	0.30
30-	0.001	0.28
35-	0.001	0.26
40-	0.001	0.24
45-	0.002	0.22
50-	0.002	0.20
55-	0.002	0.18
60-	0.003	0.16
65-	0.003	0.14
70-	0.005	0.12
75-	0.011	0.10
80-	0.021	0.08
85-	0.035	0.06
90-	0.054	0.04
95-	0.076	0.02
100-	0.103	0.00

9.21 Conclusions

The Cumpston model simulates the movement of whole households between dwellings in each of the 57 statistical divisions. It also simulates the movement of individuals out of existing households, to set up new households or join existing ones. Immigrants are modelled as setting up new households or joining existing ones.

Confidentiality restrictions and survey non-responses reduced the reliability of the HILDA data for assessing exit rates, type changes and follower probabilities, and some large adjustments were needed to exit rate assumptions. With these adjustments, the model gave type/age distributions at 30/6/10 broadly similar to those projected by ABS.

The Cumpston model uses sampling with loaded probabilities for births, deaths, emigration, exits and moves. The results in 9.13 show that this technique gives reasonable projections of person types by age group.

The results in 9.19 show that immediate best of n matching closely replicates the means and standard deviations of age differences for partners, parent/child combinations, other related persons, unrelated persons and group members.

The runtimes in 7.21 show that household exits and moves only take about 9 seconds in a 50 year simulation. This low time reflects efficiency gains from both sampling with loaded probabilities and best of n matching.

10. EDUCATION AND TRAINING

Australia has been standardizing its school education systems across the states and territories, and most students now have a year of pre-school, 7 years of primary school, and up to 5 years of secondary school.

Technical and further education (TAFE) colleges have traditionally provided certificates for tradesmen, but are now also providing some diplomas, and large numbers of non-vocational courses.

Australian universities provide a wide range of undergraduate and postgraduate courses, and have recently experienced large increases in the numbers of students from overseas.

The Cumpston model simulates 13 levels of school education, 3 of technical and further education (TAFE) and 3 of university education. Commencement, completion and exit assumptions are largely based on administrative statistics, supplemented by TAFE transition probabilities.

Baseline unit records were derived from the 2001 census sample, adjusting for the large over-statements of the youngest and oldest school students, and of university students, evident In the census. Allocations between school grades, and between TAFE and university levels, were based on administrative data.

10.1 Structure of Cumpston model

The levels of education selected for modelling are

- Grades 0 to 12 of the primary and secondary education systems (where grade 0 is commonly called pre-school)
- Technical and further education (TAFE), separately for certificate III, certificate IV and diploma or higher
- University undergraduate, graduate certificate and other postgraduate.

To simplify modelling, a number of restrictions are imposed

- All students other than immigrants enter school at grade 0
- Progression is upwards annually to grade 12, but each year can be repeated, and promotion by an additional grade can sometimes occur
- Students who do not exit year 12 are assumed to repeat that year
- No returns to primary or secondary education occur by persons who have left school
- Technical and further education can begin after the completion of year 10 or higher
- University undergraduate study cannot begin until after the completion of year 12 or the completion of a TAFE course at certificate III or higher
- University postgraduate study cannot begin until after the completion of a university undergraduate course
- Entry to TAFE and undergraduate study can occur at any age under 70, and entry to postgraduate study at any age under 65
- All education changes are assumed to occur on January 1.

10.2 Entry rates to primary education

All children (other than immigrants) are assumed to enter grade 0 on January 1, with probabilities of entry determined by their age at June 30 after entry. These probabilities were chosen so as to reproduce the age-profile of students In grade 0 at 30/6/09 (ABS 2010a). Although similar data for 2001 show that about 30% of new students entered in grade 1 rather than grade 0, entry at grade 0 appeared to be universal by 2009. Entry rate assumptions are in table 10.1.

Table 10.1 Assumed probabilities of entry to school, if not at school

Age	4	5	6	7
Probability	0.021	0.805	0.992	1.000

10.3 Repeat and promotion rates

Table 10.2 shows that 78.8% of grade 0 students in 2009 were at the median age of 5, with 2.1% below the median age and 19.1% above. Grade 12 showed a wider age spread, with 13% below the median age of 17, and 23.3% above it.

Table 10.2 Estimated repeat and promotion rates

Grade	Below median age	At median age	Above median age
0	0.021	0.788	0.191
12	0.130	0.637	0.233
Average		0.713	

An approximate estimate of the annual repeat rate was obtained as

Increase from grade 0 to 12 in proportion below median age	.109
divided by average proportion at median age	.713
divided by number of years from grade 0 to 12	12
estimated repeat rate pa	.013

An annual promotion rate of 0.5% was estimated in the same way as the repeat rate, using the percentages above the median age. Visual inspection of the age/grade distribution of school students in 2009 suggests that most promotions occur from grade 1 to 3, and from grade 6 to 8.

Although it is commonly thought that repeats are undesirable, some do occur in the lowest and highest grades. Repeat rates for grades 10, 11 & 12 were estimated as 1%, 2% and 5% by Gorgens & Ryan (2005 p12). Better estimates of repeat and promotion rates would require longitudinal data on school students.

10.4 Exit rates from schooling

Participation rates were calculated from the numbers of full-time and part-time school students at each age in 2009 (ABS 2010a) and estimated resident persons at 30/6/09 (ABS 2010b). Exit rates were then derived from these participation rates. All exits are assumed to occur on 31 December, with probabilities of exit determined by their age at June 30 after exit. Small numbers of exits occur before the legal minimum of 15. The assumed exit rates are in table 10.3.

Table 10.3 Assumed probabilities of exit from schooling

Age at 30/6	Male	Female
12	0.004	0.004
13	0.004	0.004
14	0.004	0.004
15	0.036	0.025
16	0.122	0.083
17	0.257	0.223
18	0.747	0.782
19	0.871	0.872
20+	0.712	0.671

10.5 Assumed transitions from school to tertiary education

80% of persons exiting years 10 and 11 were assumed to enter TAFE, with the balance not entering further study immediately. Similarly, 80% of persons exiting year 12 were assumed to enter university, with the balance not entering further study immediately. These very approximate assumptions were obtained by comparing the numbers leaving school at each age with the numbers entering TAFE and university (NCVER 2011 and DEEWR 2011a). TAFE entrants were assumed split between certificate II, certificate IV and higher courses in the proportions reported for TAFE enrolments in 2009 (NCVER 2011).

10.6 Entry rates to technical and further education

The assumed probailities of entry in table 10.4 are for those not currently studying but eligible to enter, with year 10 or higher completion for TAFE, year 12 or TAFE completion for undergraduate, and undergraduate completion for postgraduate. As a result of the high study levels at age 19 and below, it was difficult to estimate the eligible numbers, and the assumptions at these ages are particularly uncertain. For example, the estimated probability of 0.275 for persons aged 19 entering “other postgraduate” are estimates for the very small group of advancd students who have already completed an undergraduate degree.

Numbers of TAFE students, subdivided by the level of their major enrolment, studying at different levels of technical and further education institutions in 2009 were available from NCVER (2011 table 10). The age distribution of all TAFE students in 2008 and 2009 (NCVER 2011 table 13) was assumed to apply to each enrolment level, and used to estimate the numbers of students at certificate III level or higher in each age. Similarly, the numbers of students completing TAFE courses in 2008 at each age were estimated from NCVER table 13. Completion and exit probabilities estimated by NCVER (Roberts 2010) were used to derive from the 2008 completions approximate estimates of 2008 exits. The numbers of new students in 2009 were then derived as the increase in student numbers from 2008 to 2009, plus completions and exits in

2008, allowing for the one-year increase in age from 2008 to 2009. University commencements were available from DEEWR 2010a.

Table 10.4 Assumed probabilities of entry to TAFE and university

Age	TAFE	Under-graduate	Graduate certi-ficate	Other post-graduate
15-	0.291	0.486	0.063	0.065
19-	0.291	0.193	0.218	0.275
20-	0.194	0.102	0.101	0.144
21-	0.162	0.072	0.061	0.083
22-	0.128	0.048	0.050	0.054
23-	0.105	0.034	0.040	0.043
24-	0.089	0.025	0.029	0.035
25-	0.074	0.016	0.024	0.027
30-	0.050	0.011	0.015	0.017
35-	0.046	0.007	0.017	0.017
40-	0.049	0.007	0.012	0.012
45-	0.041	0.005	0.008	0.008
50-	0.030	0.003	0.006	0.006
55-	0.019	0.002	0.002	0.002
60-	0.013	0.001	0.002	0.003
65-	0.007	0.001	0.000	0.000

10.7 Assumed TAFE completion and exit rates

The completion and exit probabilities in table 10.5 were estimated by Roberts (2010) as described by Mark & Karmel (2010), using Australian unit records from 2005 and 2006. They are here assumed to apply from 2002 on.

Table 10.5 Completion and exit rates for TAFE courses and years

Level	Event	Year	Male 15-19	Male 20-24	Male 25-	Female 15-19	Female 20-24	Female 25-
Bachelor	Completion	1	0.000	0.027	0.069	0.000	0.043	0.109
Diploma	Completion	1	0.123	0.163	0.139	0.121	0.161	0.137
Cert IV	Completion	1	0.206	0.192	0.214	0.250	0.232	0.260
Cert III	Completion	1	0.094	0.137	0.148	0.168	0.245	0.262
Bachelor	Completion	2	0.000	0.207	0.266	0.000	0.283	0.364
Diploma	Completion	2	0.356	0.333	0.259	0.434	0.407	0.315
Cert IV	Completion	2	0.258	0.271	0.291	0.316	0.333	0.357
Cert III	Completion	2	0.163	0.261	0.280	0.249	0.397	0.426
Bachelor	Exit	1	0.381	0.449	0.403	0.415	0.489	0.439
Diploma	Exit	1	0.525	0.525	0.535	0.477	0.477	0.485
Cert IV	Exit	1	0.640	0.607	0.577	0.572	0.543	0.517
Cert III	Exit	1	0.410	0.466	0.542	0.414	0.470	0.548
Bachelor	Exit	2	0.498	0.480	0.322	0.350	0.338	0.226
Diploma	Exit	2	0.425	0.457	0.488	0.373	0.401	0.428
Cert IV	Exit	2	0.571	0.546	0.550	0.503	0.482	0.484
Cert III	Exit	2	0.252	0.250	0.344	0.330	0.328	0.450

10.8 Assumed university completion and exit rates

University completion rates for all students were estimated by comparing the numbers of all completions in 2009 (DEEWR 2010b table 5) with the numbers of all students (DEEWR 2010a table 2.1). The resulting rates were then slightly adjusted to give the correct overall numbers of domestic completions. University exit rates were approximately estimated by comparing the numbers of all students in 2008 and 2009 with completions in 2008 and new students in 2009. Assumed rates are in table 10.6.

Table 10.6 Assumed probabilities of completion and exit

Age	Complete Under-graduate	Complete Graduate certificate	Complete Other post-graduate	Exit Under-graduate	Exit Graduate certificate	Exit Other post-graduate
-18	0.021	0.252	0.091	0.114	0.158	0.078
19	0.082	0.312	0.091	0.165	0.158	0.078
20	0.209	0.495	0.103	0.170	0.264	0.078
21	0.332	0.513	0.144	0.170	0.264	0.078
22	0.389	0.532	0.203	0.170	0.264	0.078
23	0.376	0.534	0.266	0.170	0.264	0.078
24	0.334	0.546	0.298	0.170	0.264	0.078
25-	0.264	0.470	0.268	0.158	0.350	0.135
30-	0.203	0.406	0.223	0.112	0.317	0.135
40-	0.194	0.418	0.203	0.126	0.307	0.135
50-	0.193	0.402	0.185	0.083	0.296	0.135
60-	0.173	0.326	0.170	0.063	0.238	0.135

10.9 Baseline data

The “type of educational institution attending” field in the 2001 Household Sample File was used to determine type of study at 30/6/01. Persons with “not stated” type were allocated to other HSF types, including not studying, in proportion to the stated values for their age and sex. Persons with “other” type were allocated to the seven specific HSF types in proportion to the values for their age and sex. Persons with HSF type “pre-school” were assumed to be in study level 0. Persons in infants/primary or secondary school were allocated to study levels 1 to 12, based on their age and sex, using the age/sex distribution of full-time and part-time school students in 2001 (ABS 2010a).

The derived numbers of students in study levels 0, 1 and 12 proved significantly higher than the known numbers of students in 2001 (ABS 2010a), so 38% of those supposedly in levels 0 and 1 were assumed not be studying, and 14% of those in level 12 were assumed not to be studying. It is not clear why the estimates derived from 2001 census data were higher than the official records. One possibility is that many children reported as being in pre-school were in fact in child care. The year 12 excess may reflect a tendency to exaggerate education level.

The numbers of persons estimated from the census as attending TAFE were only 78% of NCVER (2011) enrolments for certificate III or higher courses in 2001. The reasons for this shortfall are not known, and no attempt was made to correct for this initial shortfall in TAFE students in the simulations. Students were allocated between certificate III, certificate IV and higher courses using the distribution of certificate III and higher students reported by NCVER (2011) for 2001.

Persons described in the census as university students were allocated between undergraduate, graduate certificate and other postgraduate courses using the age distribution reported by DEEWR (2010a) for 2001. The resulting numbers were significantly higher than DEEWR totals, and 26% of those in undergraduate courses, and 45% of those in postgraduate courses, were assumed not to be studying. Census respondents appear to exaggerate their involvement in university study.

10.10 Reasonableness checks

Given the indirect derivations of some of the education assumptions, and the major uncertainties in the baseline data, it was important to check the reasonableness of the projection results. Actual numbers of school students in table 10.7 are from ABS (2011a table NSSC 40a), of TAFE students are from NCVER (2011 table 10), and of numbers of domestic university students are from DEEWR (2011a table 2.6). The simulated numbers of students are the averages of 3 runs, and are broadly comparable with actual numbers.

Table 10.7 Simulated and actual numbers of students at each level at 30/6/10

Study level	Simulated m	Actual m	Simulated/ actual
Pre-school	0.270	0.282	0.96
Grade 1	0.285	0.269	1.06
Grade 2	0.288	0.250	1.15
Grade 3	0.295	0.271	1.09
Grade 4	0.289	0.275	1.05
Grade 5	0.286	0.274	1.04
Grade 6	0.274	0.274	1.00
Grade 7	0.237	0.266	0.89
Grade 8	0.265	0.280	0.95
Grade 9	0.281	0.283	0.99
Grade 10	0.257	0.280	0.92
Grade 11	0.233	0.261	0.89
Grade 12	0.235	0.220	1.07
TAFE III	0.596	0.553	1.08
TAFE IV	0.211	0.254	0.83
TAFE higher	0.218	0.233	0.93
Undergraduate	0.600	0.605	0.99
Graduate certificate	0.062	0.069	0.89
Other postgraduate	0.146	0.139	1.05
Total	5.327	5.339	1.00

Actual numbers of TAFE completions in table 10.8 are from NCVER (2011 table 14), extrapolating from 2008 and 2009 values. Actual numbers of domestic university completions are from DEEWR (2011b table 9), interpolating between 2009 and 2010 numbers. Simulated numbers are the averages of 3 runs.

Simulated TAFE course completions in table 10.8 are about 24% below actual completions. TAFE completion numbers have increased faster than student numbers in recent years, so the completion rates assumed here, based on 2005 and 2006 experience, may be too low. Simulated university completions are about 10% above actual completions.

Table 10.8 Simulated and actual numbers of course completions in 09-10

Study level	Simulated m	Actual m	Simulated/ actual
TAFE III	0.135	0.170	0.79
TAFE IV	0.054	0.080	0.67
TAFE higher	0.045	0.058	0.77
Undergraduate	0.131	0.111	1.18
Graduate certificate	0.027	0.030	0.90
Other postgraduate	0.032	0.031	1.03

10.11 Education in six other national models

Payne, Percival & Harding (2008) gave an initial proposal for how Australia's education system would be modelled in APPSIM. They proposed full-time education be compulsory up to year 10, with an age cut-off for returning to education of 30. Adults were only to be allowed to return to a higher course than the highest qualification they had completed. APPSIM assumes 10% of primary students repeat each year, and much lower levels of secondary students repeat.

DYNACAN modelled up to 12 years of school education, then provided an "instant upgrade" for selected 32-year olds to reflect the impact of later-life education (Morrison 2008 p16).

The MOSART model is extensively used to project labour supply by educational attainment, and models 7 fields of upper secondary education, 9 of lower tertiary education and 10 of upper tertiary education (Gunnes 2009). No return to education appears to be possible.

SESIM III assumes all individuals start gymnasium, and models their exits, including to university study and the labour force. Transitions from the labour force to adult gymnasium, and from adult gymnasium to university, are also modelled. Exits from university study are modelled with and without a degree (Flood et al 2005 p32-33).

DYNASIM3 models education at grade school, high school, college, and graduate school. Exit is possible from grade 8 on. Probabilities of progression are based in part on race, sex and parental education (Favreault & Smith 2004 p10). The authors note that this structure is consistent with sociological literature on status attainment, which has consistently found that parents' status is one of the strongest predictors of a child's success in school.

Pensim2 randomly assigns a level of education to all individuals at age 16, and the age of leaving education is calculated from this level (Emmerson, Reed & Shephard 2004 p26).

INAHSIM does not model education (Inagaki 2011 p9), but education is allowed for in the z-score, representing a person's ability to earn money.

10.12 Conclusions

Simulated numbers of students for the 13 school levels, 3 TAFE levels and 3 university level are broadly comparable with actual numbers at 30/6/10. Simulated numbers of TAFE completions are 24% below actual completions in 09-10, and simulated

university completions 10% above. Adjustments are needed to more closely match recent experience.

If adjusted, the simulated results should be broadly adequate for use elsewhere in the model, for example, in modeling fertility, partnership, employment, occupation and earnings.

More detailed models may be needed for some purposes, such as estimating student loan debts or new graduates in particular fields. As well as subdividing projections by field of study, it may be desirable to model completion and exit rates on a duration dependent basis, taking into account the years of full-time study required to complete particular qualifications.

11. EMPLOYMENT, OCCUPATION AND EARNINGS

Employment as an employee, and as self-employed, are separately modelled using logistic probability models. A person can be both an employee and self-employed at the same time. An occupation is selected for persons entering employment, and changes of occupation are modelled for persons remaining in employment. As persons can enter or re-enter employment at any age, unemployment and retirement are not modelled,

Wages for employees are modelled using logarithmic random effects models. Business incomes for the self-employed, which can be negative or positive, are modelled using random effects models without transformations.

Other national models use a range of techniques to simulate workforce participation and earnings. DYNASIM, DYNACAN, INHASIM, SESIM allow for permanent individual earning differences, using random effects models or permanent “luck” variables. Occupations and negative incomes are not commonly modelled.

11.1 Structure of Cumpston model

Each simulation period, the following are simulated for each person

- Being an employee
- Being self-employed
- If not previously employed and now employed, occupation
- If previously employed and now employed, any change of occupation
- If an employee, wages in the period
- If self-employed, business income in the period

All the logistic probability and random effects models are specified separately for males and females. Separate models are used for the probabilities of being an employee or self-employed, depending on whether the person was previously employed. If a person is an employee and self-employed at the same time, the same occupation is assumed for both.

No distinction is made between unemployment and persons choosing not to be in the labour force. Instead, any person who is not an employee or self-employed is classed as not employed, with probabilities of returning to employment in each simulation period. No distinction is made between part-time and full-time employees. The effects of full-time and part-time study are allowed for with an explanatory variable where significant. Depending on model applications, a more complex structure might be appropriate.

At present, employment, occupation, wages and business incomes are simulated on an annual cycle, with no use being made of sampling with loaded probabilities. Multithreading should be a useful way to reduce runtimes and allow shorter simulation cycles.

Each employee and self-employed person is assigned one of the occupations in table 11.1. An occupation is assigned on entry to employment, and transitions between occupations are modelled.

Table 11.1 Occupation classifications used in 2001 census

Occupation	Occupation description
1	managers & administrators
2	professionals
3	associate professionals
4	tradespersons & related workers
5	advanced clerical & service workers
6	intermediate clerical, sale s & service workers
7	intermediate production & transport workers
8	elementary clerical, sales & service workers
9	labourers & related workers

11.2 Probabilities of entry to employment

Table 11.2 shows the coefficients of logistic models of entry to employment, given that a person was neither an employee nor self-employed in the previous financial year. These were derived from HILDA waves 1 to 6. The second and third columns are the models for entry as an employee, separately for males and females. The last two columns are the models for entry to self-employment. Entry as both an employee and self-employed can occur in the same year. The coefficients shown were obtained by stepwise regression, dropping those not significant at the 5% level.

Table 11.2 Logistic models of entry to employment, if not employed last year

Variable	Employee Male	Employee Female	Business Male	Business Female
age		0.049	0.086	0.129
agesquare	-0.0010	-0.0014	-0.0013	-0.0017
full-time student	-0.704		-1.151	
group	1.114			
loneparent	-0.728	-0.289		
loneperson	0.376			
partner	0.666		0.621	0.910
postgraduate	1.000	1.088		0.983
tafe	0.279	0.453	0.312	
children 0-4	-0.351	-0.729		
children 5-14		-0.165		0.201
undergraduate	0.738	0.661	0.686	0.693
constant	0.241	-0.932	-4.584	-6.367
N	7753	12333	7753	12294
LLR	0.273	0.235	0.081	0.104

Probabilities of becoming an employee reduce with age, while probabilities of becoming self-employed peak at about age 34 for males and 39 for females. Postgraduate qualifications increase probabilities of entry to employment a little more than undergraduate qualifications. Having young children in the household reduce the probabilities of entry to employment, particularly for females.

11.3 Probabilities of wages or business if employed last year

Table 11.3 shows the coefficients of logistic models of being an employee or self-employed, given that a person was an employee or self-employed in the previous financial year. These were derived from HILDA waves 1 to 6. The second and third columns are the models for being an employee, separately for males and females. The last two columns are the models for being self-employed. As noted, persons can be both employees and self-employed in the same year. The coefficients shown were obtained by stepwise regression, dropping those not significant at the 5% level. The age coefficient for females in business was not significant at the 5% level, but was retained for completeness. "Occlag1" is a dummy variable for persons with occupation 1 in the prior year.

Table 11.3 Logistic models of employment, if employed last year

Variable	Employee Male	Employee Female	Business Male	Business Female
age	0.067	0.110	0.087	0.036
agesquare	-0.0009	-0.0014	-0.0008	-0.0002
business last year	-2.121	-1.876	3.984	3.978
occlag1	0.342		0.421	0.404
occlag2	0.905	0.865		
occlag3	0.791	0.476		
occlag4	0.352	0.441		
occlag5	1.235	0.498		0.281
occlag6	0.971	0.601	-0.489	-0.359
occlag7	0.587		-0.589	
occlag8	0.571	0.625		-0.362
partner	0.151	0.202	0.246	0.507
post	0.303	0.239		
tafe	0.110			
children 0-4		-0.139		0.159
children 5-14	-0.064			0.114
under	0.199	0.171		
employee last year	1.964	2.127	-1.320	-1.027
constant	-0.909	-2.015	-3.715	-3.435
adjust	-0.100	-0.100	-0.100	-0.100
N	19482	17586	19482	17586
LLR	0.379	0.315	0.576	0.537

Probabilities of remaining in employment as an employee appear to peak at about 38 for males and 40 for females, and in business at 53 for males. Persons who were self-employed last year have much higher probabilities of being self-employed this year, and much lower probabilities of being employees. Similarly, persons who were employees last year have higher probabilities of being employees, and lower probabilities of being self-employed.

11.4 Probabilities of entry to different occupations

Table 11.4 shows logistic models of probabilities of entry into each occupation, derived from HILDA waves 1 to 6.

Table 11.4 Logistic models of probabilities of entry to each occupation

Variable	occ1	occ2	occ3	occ4	occ5	occ6	occ7	occ8	occ9
age	0.030			-0.016		0.057		-0.116	0.068
agesquare						-0.001		0.001	-0.001
full-time student								0.368	
partner			0.767		1.673				
post		1.844							
sex				-1.959	1.587		-1.543	0.578	-0.556
tcr04					0.520	1.257			-0.480
tcr514							0.331		
under		0.926					-1.069		-0.865
constant	-4.832	-3.106	-3.716	0.215	-8.424	-5.028	-1.118	-1.002	-2.383
N	2089	2701	2701	2633	2633	2701	2633	2701	2633
LLR	0.027	0.036	0.016	0.104	0.118	0.053	0.067	0.057	0.027

Having a postgraduate or undergraduate degree greatly increases the chances of becoming a professional (occupation 2). Females are less likely to enter occupations 4, 7 and 9, and more likely to enter occupations 5 and 8. After persons were simulated to enter employment using the probability models in table 11.2, the probabilities of entering each occupation were calculated using the 9 models in table 11.4 and an occupation randomly selected on a probability weighted basis.

11.5 Logistic models of occupation change

Table 11.5 shows logistic models of probabilities of changing occupation, derived from HILDA waves 1 to 6.

Table 11.5 Logistic model of probability of changing occupation

Coefficient	Male	Female
age	-0.017	
agesquare		-0.0002
full-time student	-0.249	-0.379
group		0.480
loneperson		0.158
occlag1	-0.227	0.407
occlag2	-0.773	-1.118
occlag3		0.259
occlag4	-0.986	0.222
occlag5	0.755	0.222
occlag6	0.151	-0.277
occlag7	-0.492	0.354
post		0.319
children 0-4		-0.217
under		0.186
constant	0.145	-0.521
N	19486	17639
LLR	0.037	0.045

Professionals and male tradespersons have unusually low probabilities of changing occupation.

11.6 Logistic models of new occupation

Table 11.6 shows logistic models of the probability of entering a new occupation, given that a change of occupation has occurred. These were derived from HILDA waves 1 to 6.

Table 11.6 Logistic models of probability of entering an occupation, given a change

Coefficient	occ1	occ2	occ3	occ4	occ5	occ6	occ7	occ8	occ9
age	0.096		0.095		0.078	0.048	0.090	-0.077	
agesquare	-0.001	0.000	-0.001	0.000	-0.001	-0.001	-0.001	0.001	
full-time student	-1.156		-0.636	-0.386			-0.565	0.780	
group		0.748							
loneperson		0.210	0.278		-0.344				
occlag1		2.408	1.429	-1.078	1.505	-0.373	-1.653	-1.292	0.848
occlag2	1.113		1.529	-1.295	1.378		-1.758		-0.770
occlag3	0.478	2.073		-0.748	1.766	0.556	-1.423		
occlag4		1.165	1.135			-0.668	-0.663		1.839
occlag5		1.524	1.248	-1.658		0.740	-1.802	-0.508	-0.760
occlag6		1.949	1.418	-1.757	2.136		-1.005	0.728	
occlag7	-0.412	0.851	0.324	-0.483				0.635	1.705
occlag8	-0.984	1.155	0.824	-1.106	0.766	0.810	-0.588		
partner	0.240		0.244					-0.248	-0.317
post	1.309	1.798		-2.164	-1.432	-0.777	-1.888	-1.325	-1.846
sex	-0.630			-1.748	1.815	0.934	-1.686	0.423	
tafe			0.131	0.742			-0.190	-0.237	-0.304
under	0.602	1.028		-0.631		-0.220	-0.809	-0.581	-0.697
constant	-3.992	-3.612	-4.296	1.048	-9.069	-3.564	-0.619	-1.271	-2.520
N	10241	9823	9579	10493	10873	9389	10560	10060	11395
LLR	0.136	0.133	0.058	0.181	0.150	0.093	0.151	0.107	0.124

The logistic models in table 11.6 are intended for use in simulating new occupations, for those persons where a change of occupation has already been simulated using the model in table 11.5. The logistic probabilities are worked out for each possible new occupation, and an occupation randomly selected on a probability-weighted basis.

11.7 Random effects models of log(wages)

Table 11.7 shows three different types of random effects models, fitted separately to males and females in HILD waves 1 to 6.

Table 11.7 Random effects models of log(wages)

Coefficient	AR(0) Male	AR(0) Female	AR(1) Male	AR(1) Female	Lagged Male	Lagged Female
age	0.167	0.149	0.160	0.144	0.157	0.140
agesquare	-0.0019	-0.0018	-0.0019	-0.0018	-0.0018	-0.0017
busflag	-0.166	-0.129	-0.150	-0.113	-0.157	-0.119
full-time student	-0.576	-0.541	-0.533	-0.506	-0.574	-0.544
group	0.228		0.190		0.234	
loneparent	0.189	0.101	0.190	0.111	0.190	0.111
loneperson	0.213	0.182	0.213	0.182	0.212	0.182
occ1	0.429	0.678	0.457	0.706	0.426	0.679
occ2	0.386	0.603	0.400	0.611	0.379	0.598
occ3	0.340	0.495	0.350	0.503	0.334	0.493
occ4	0.289	0.262	0.288	0.249	0.282	0.265
occ5	0.285	0.433	0.294	0.455	0.280	0.434
occ6	0.221	0.302	0.216	0.302	0.216	0.304
occ7	0.193	0.210	0.203	0.218	0.191	0.207
occ8	0.062	0.107	0.059	0.108	0.051	0.117
partner	0.292	0.129	0.307	0.126	0.290	0.124
post	0.329	0.322	0.319	0.317	0.312	0.307
children 0-4		-0.426		-0.390		-0.414
children 15-24		-0.078		-0.077		-0.074
children 5-14	-0.044	-0.209	-0.040	-0.199	-0.040	-0.199
under	0.142	0.210	0.142	0.210	0.132	0.198
wage1	-0.532	-0.777	-0.580	-0.834	-0.239	-0.513
wage2	-0.213	-0.362	-0.258	-0.408	-0.192	-0.336
wage3	-0.109	-0.221	-0.136	-0.256	-0.105	-0.217
logwageslag					0.031	0.003
constant	6.636	6.757	6.753	6.844	6.542	6.658
observations	16994	16364	16994	16364	16994	16364
persons	5107	4944	5107	4944	5107	4944
overall Rsquare	0.436	0.397	0.440	0.400	0.458	0.418
sigma_u	0.640	0.694	0.529	0.561	0.610	0.657
sigma_e	0.475	0.552	0.511	0.588	0.474	0.552
autocorrelation			0.301	0.286		

Table 11.7 gives the coefficients of random effects regression models fitted to the log of wages. Only records with wages greater than 0, and known occupation, were used. To allow for entry and wage lag effects, the models were fitted to the second and subsequent records for each person. For example, there were 22,101 records for 5,107 males, and omitting the first record for each male left 16,994 for model-fitting.

Initially, AR(0) models were fitted, with no regression in the disturbance terms. As suggested by O'Donoghue, Leach and Hynes (2007), trials were also made with AR(1) models, where the disturbance terms were first-order autoregressive, and with lagged models, where the log of the wages in the preceding year was used as one of the explanatory variables. The autocorrelations found with the AR(1) models were low, and the coefficients for lagged wages were also low. Given that each of the models

gave similar overall Rsquare results, the AR(0) models seem preferable on the grounds of simplicity.

The coefficients for age and agesquare suggest that wages peak at about 43 for males and 41 for females. As expected, persons in full-time education earn much less.

The model being fitted is

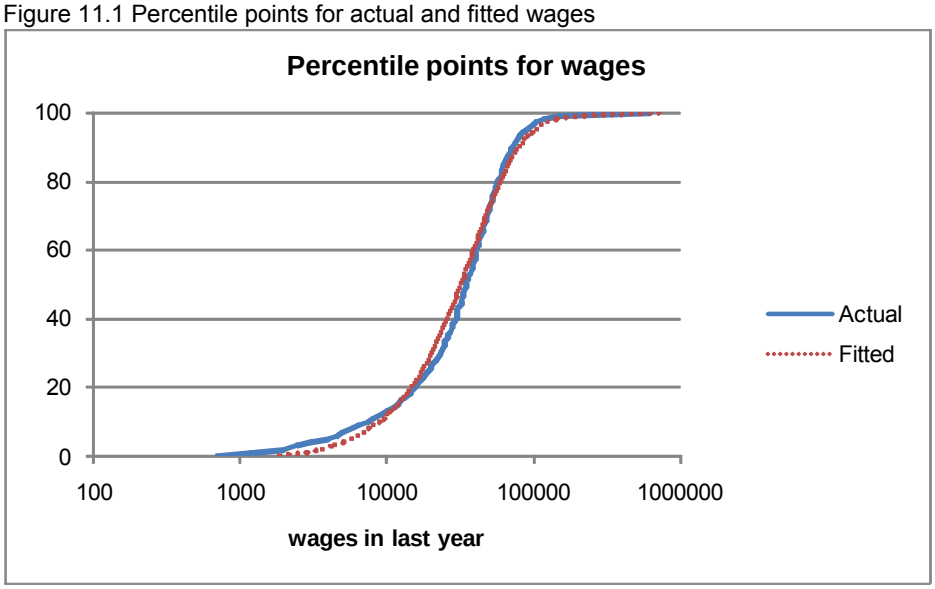
$$y_{it} = \alpha + x_{it}\beta + \mu_i + \varepsilon_{it}$$

where y_{it} is the observed value for the i^{th} individual in the t^{th} period,
α is a constant
x_{it} is a matrix of independent variables for individual i at time t
β is a vector of regression coefficients
μ_i is the fixed offset for individual i
and ε_{it} is a random error term.

Models of this form can be fitted in several ways, including random effects, between-effects, fixed effects, and population averaged (STATA 2007 p390-413). A random effects model was fitted here, as coefficients were required for static variables such as sex, and an estimate was required of the standard deviation of individual offsets.

Figure 11.1 shows that a reasonable fit to actual wages was obtained, except at the lowest wages. The standard deviation of the fitted wages is 7% higher than the standard deviation for actual wages. The fitted wages are derived from the predicted log values for each data point, adding a randomly generated normal variable with zero mean and standard deviation equal to the standard deviation of the random error term, and exponentiating.

Figure 11.1 Percentile points for actual and fitted wages



11.8 Random effects model of business income

Table 11.8 shows random effects models of business income derived from HILDA waves 1 to 6. As business income can be negative, a model of log(business income) could not be fitted.

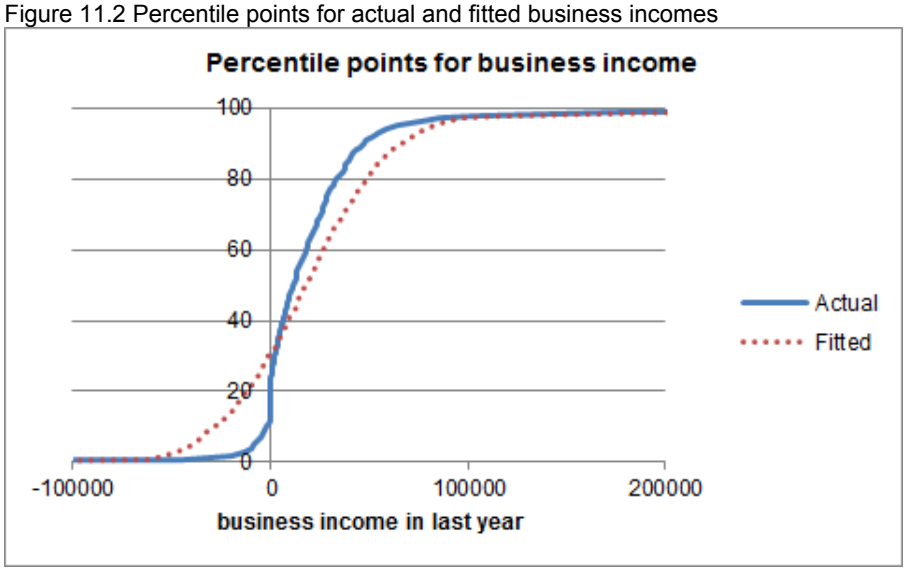
Table 11.8 Random effects model of business income

Coefficient	Male	Female
age	1074	550
agesquare	-13	-5
bus1	-6164	-4577
employee	-14846	-5685
constant	4859	3068
observations	3001	1956
persons	1266	879
overall Rsquare	0.044	0.041
sigma_u	31356	29330
sigma_e	30918	12409

Table 11.8 gives the coefficients of a random effects regression model fitted to own business income from unincorporated business.

Figure 11.2 shows that a poor fit to actual business income was obtained. The standard deviation of the fitted values is about 6% less than that of the actual values. The fitted salaries are derived from the predicted values for each data point, adding a randomly generated normal variable with zero mean and standard deviation equal to the standard deviation of the random error term

Figure 11.2 Percentile points for actual and fitted business incomes



11.9 Baseline data

Estimates of wages and business income were derived from the 2001 1% sample file with some difficulty, as the census only asks for the total of income from all sources. Persons with negative income were assumed to be receiving only business income. For employed persons with positive incomes, regressions were used to estimate the probabilities of receipt of wages, business income and other income. The receipt of each of these forms of income was simulated, and the total income shared between

the relevant types in proportion to the probabilities of receipt. Person with "not stated" occupations were randomly allocated occupations based on the occupation distribution for their labour force status.

11.10 Value of occupation as an explanatory variable

The models for employment if previously employed, and for real wages, were rerun for males, omitting all the occupation explanatory variables. As table 11.9 shows, occupation modestly increases the R square value, or the log-likelihood reduction

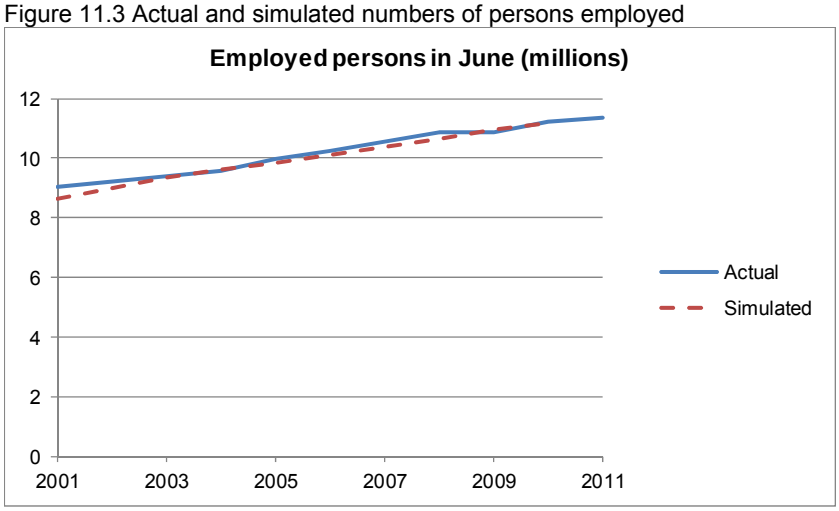
Table 11.9 Value of occupation as an explanatory variable

Model	Rsquare or LLR without occupation	Rsquare or LLR with occupation	Increase due to occupation as %
probability employee if employed last year	0.3716	0.379	2%
probability self-employed if employed last year	0.5728	0.5764	1%
random effects model for log wages	0.4017	0.4312	7%

11.11 Checks on first 10 projection years

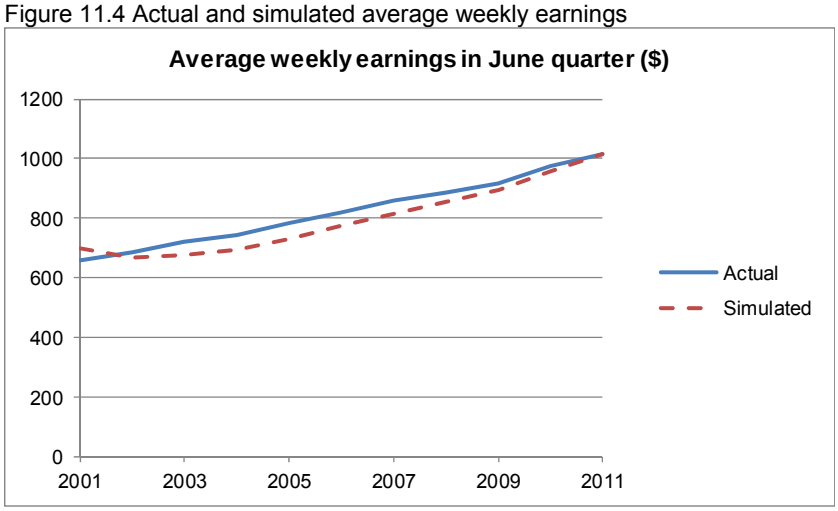
The simulated numbers of employed persons in figure 11.3 were obtained using the logistic models in tables 11.2 and 11.3, after deducting 0.1 from the regression constants In the second and third columns, and 0.8 from the regression constants in the fourth and fifth columns. These deductions were made so as to get an approximate balance between actual and simulated numbers in June 2011. Actual values are from ABS 2011e table 7.

Figure 11.3 Actual and simulated numbers of persons employed



The simulated values of average weekly earnings in figure 11.4 were obtained in June 2001 values using the random effects models in the second and third columns of table 11.12. Allowance was then made for actual wage inflation since June 2001, plus an additional 1.2% pa inflation intended to approximately replicate average weekly earnings in June 2011.

Figure 11.4 Actual and simulated average weekly earnings



11.12 Simulated and actual occupations in June 2011

Table 11.10 shows the scaled-up numbers of persons employed in each occupation in the baseline data, and the numbers simulated at 30/6/11.

Table 11.10 Simulated occupations in 2001 and 2011 (2001 classification)

Occupation	2001 M	2001 F	2011 M	2011 F	Increase M	Increase F
Managers & administrators	0.558	0.226	0.683	0.349	22%	54%
Professionals	0.754	0.840	0.959	1.392	27%	66%
Associate professionals	0.588	0.443	0.831	0.593	41%	34%
Tradespersons & related workers	0.985	0.125	1.182	0.153	20%	22%
Advanced clerical & service	0.034	0.293	0.049	0.401	46%	37%
Intermediate clerical, sales & service	0.415	1.045	0.548	1.526	32%	46%
Intermediate production & transport	0.621	0.098	0.772	0.106	24%	8%
Elementary clerical, sales & service	0.295	0.543	0.362	0.632	23%	16%
Labourers & related workers	0.472	0.288	0.596	0.314	26%	9%
Total	4.720	3.901	5.982	5.464	27%	40%

Table 11.11 shows estimated numbers of persons in each occupation, derived from surveys in August of each year (ABS 2011e). The occupational classification system was changed between 2001 and 2011, so direct comparisons are not possible.

Table 11.11 Actual numbers in each occupation in 2001 and 2011 (2011 classification)

Occupation	2001 M	2001 F	2011 M	2011 F	Increase M	Increase F
Managers	0.760	0.351	0.951	0.521	25%	48%
Professionals	0.918	0.845	1.169	1.277	27%	51%
Technicians & trade workers	1.177	0.175	1.404	0.210	19%	20%
Community & trade services	0.242	0.509	0.351	0.753	45%	48%
Clerical & administrative	0.351	1.152	0.407	1.301	16%	13%
Sales workers	0.340	0.565	0.414	0.648	22%	15%
Machinery operators & drivers	0.570	0.062	0.704	0.071	24%	15%
Labourers	0.661	0.366	0.764	0.398	16%	9%
Total	5.020	4.024	6.165	5.179	23%	29%

To get reasonable agreement between actual and simulated growth rates in each occupation from 2001 to 2011, the adjustments in table 11.12 were made to the regression constants for entry to each occupation in table 11.6.

Table 11.12 Adjustments to regression constants for entry into each occupation

Occupation	1	2	3	4	5	6	7	8	9
Male	-2.0	-0.2	0.0	2.0	0.5	1.5	1.5	0.5	0.0
Female	-4.0	-0.5	-1.0	0.5	0.0	2.0	1.0	0.5	-0.5

11.13 Other national models

Table 11.13 gives brief details of the treatment of education in seven national models.

Table 11.13 Other national models

Model & Reference	Employment statuses	Simulation of changes in employment status	Earnings model
APPSIM Keegan 2009 p 12-13 Keegan & Thurecht 2008 p 27-39	Full-time employed, part-time employed, unemployed & not in labour force	Participation probabilities based on sex, age, education, prior employment status, health, & age of youngest child. Logit model to split employed between employees & self-employed	Not known, but likely to consist of log(wage) regressions for different groups.
DESTINIE Bonnet & Mahieu 1998 p179	Employed, unemployed, student, non-working; early retired, retired	Transition probabilities based on sex, age, school leaving age & number and age of children	Log(wage) estimated using school leaving age, duration employed and duration squared. Randomised residual correlated with its past values.
DYNACAN Morrison 1998 p17, Keegan & Thurect 2008 p25	Full time, part time, not in labour force, self-employed employee		Hours worked & earnings depend on age, sex, education, family structure & previous earnings. A permanent "luck" variable gives higher earnings.
DYNASIM Favreault & Smith 2004 p 11-12	Labour force participation for persons aged 16-80	Participation based on race, sex & age group, with random effect probit model	Hourly wage estimated using random effects log model, annual hours using tobit model including predicted wage as regressor
INHASIM Inagaki 2009a p6	Regular & irregular employment, self-employment & unemployed	Transition probabilities based on sex, age and marital status for females	Log normal by sex, age group & employment status. Lifetime score for position in earnings distribution.
SESIM III Flood et al 2005 p16-20	Employed, unemployed, retired?	Retirement based on economic utility, sex, education, household composition, labour force status of spouse, but not Age	Log random effects model, on education, blue collar, white collar, government, local government, self-employed, nationality & marital status
SVERIGE Holm et al 2002 p28-31	Employed, unemployed	Work simulated on sex, age, education level, civil status, former salary & nationality	Earnings based on age, sex, previous income, nationality

Most of the above models are using a random effects model of log(wages), or similarly a permanent “luck” variable. Most model several different employment statuses, including unemployed and not in the labour force. DYNACAN and DYNASIM model

hours worked and earnings per hour separately. The Cumpston model does not distinguish between full and part-time employment, and does not distinguish between types of non-employment. It does however allow persons to be employees and self-employed at the same time, simulates occupation changes, and by using a linear regression model for business income allows for negative business incomes (common in practice).

11.14 Conclusions

Parameters for the logistic and random effects models were derived from waves 1 to 6 of the HILDA longitudinal data. The random effects model fitted to the log of real wages gives plausible results for the effects of sex, age, occupation and years of employment. Good graphical fits are observed between the distributions of actual and fitted wages. The random effects model fitted to real self-employment incomes gives low R squared values and poor graphical fits, reflecting the diversity of self-employment.

Estimates of wages, business income and other income were derived for each person in the baseline data with some difficulty. The number of employed persons and average weekly earnings in 2001 were closely replicated. With minor adjustments to regression constants, the number of employed persons and average weekly earnings in 2011 were also replicated, and reasonable increases in the numbers in each occupation obtained.

These data analyses and test results show that logistic and random effect models derived from HILDA longitudinal data can be used to model employment, occupation wages and business incomes. Depending on the applications for the simulations, better data and more complex models might be needed. For example, if the simulations were being used to test different student loan repayment schemes, more complex models of transition from education to employment would be needed.

12. SUPERANNUATION

Tax incentives and legislative contribution requirements have seen rapid growth in Australian superannuation funds, intended to provide retirement benefits separately from social security entitlements. Increasingly, these funds provide lump sum benefits based on accumulated contributions, with longevity and investment risks borne by individuals.

Individuals can choose to keep retirement benefits within superannuation funds, drawing tax-free "allocated pensions" to meet their needs until the funds are exhausted. Their flexibility and low investment tax rates have made such pensions increasingly popular.

In 2007, the Australian government eliminated most benefit caps and taxes, while allowing large contributions into superannuation funds. Large cash inflows resulted, and reductions in allowable contributions have subsequently been made.

Using administrative data, and some limited survey data, this chapter models individual superannuation contributions, benefits and fund balances from 1/7/01 on. Substantial annuities are still being paid by legacy defined benefit schemes, and these are separately modelled.

More detailed data have recently become available for "small" superannuation funds, which are those with four members or less. Members of such funds tend to have higher contributions and account balances, and lower account closure rates. They are here modelled separately from "large" funds.

12.1 Development of superannuation in Australia

Since 1909, Australia has had mean-tested age pensions, unrelated to occupational earnings. Tax incentives and legislative contribution requirements have seen gradual growth in superannuation funds, intended to provide disability and retirement benefits separately from social security entitlements.

Prior to the 1970s, superannuation applied to a minority of employees, generally higher paid white collar staff in large corporations, employees in the finance sector, public servants and members of the Defence Force (APRA 2007 p3). In 1982-83 82% of fund members were covered by defined benefit funds, where benefits were usually defined by salary history, length of service and age.

New industrial awards including superannuation were progressively negotiated under the guidelines established by the 1986 National Wage Case. Coverage in the private sector grew from 32% in 1987 to 68% in 1991. Most of the newly covered would have been in schemes with low defined employer contributions, with benefits consisting of accumulated contributions and interest. By 2005-06 97% of members were in funds that provided either accumulation benefits or a mixture of accumulation and defined benefits.

As from 1/7/92, Australian employers were required to make superannuation contributions at least equal to 3% of wages. Exemptions applied for workers earning less than $450 a month, part-time employees under age 18 and employees aged 65 or over. From 1/7/97 the maximum age was increased to 70. The minimum employer contribution rate gradually increased, to 9% from 1/7/02.

Prior to 1/7/83, employer contributions were generally tax-deductible, and pensions and 5% of lump sum benefits were taxable. From 1/7/83, lump sums up to $50,000

were taxed at 15%, and sums above $50,000 taxed at 30%. From 1/7/88, a 15% tax applied to employer contributions, and the tax rates on lump sums were reduced by 15%. "Reasonable benefit limits" were introduced on 1/7/88, reducing the benefits available at concessional tax rates.

In 1996 a 15% surcharge was imposed on employer contributions made on behalf of individuals with an income of $85,000 or more. The surcharge rate was reduced in 03-04, and the surcharge abolished on 1/7/05.

As from 1/7/07, nearly all taxes on pensions and lump sum benefits were removed for persons aged 60 and over. Reasonable benefit limits were also removed. Employer contribution limits for tax-deductibility were simplified, with the maximum for persons aged 50 and over being reduced slightly to $100,000. As from 1/7/09, this was reduced to $50,000 (APRA 2011c).

Between 10/5/06 and 30/6/07, up to $1m in non-tax-deductible contributions could be made. As from 1/7/07, up to $150,000 of non-tax-deductible contributions could be made each year.

Total contributions to superannuation funds were $52b in 01-02, peaked abruptly at $164b in 06-07, and were $108b in 09-10. Fund assets increased from $484b at 30/6/01 to $1225b at 30/6/10 (APRA 2011b).

12.2 Structure of Cumpston model

Each employee with wages above the statutory threshold of $450 per month is assumed to receive employer superannuation contributions. The majority of employees are assumed to receive the statutory contribution of 9%, with some above and some below. Persons with wages or business income are simulated to sometimes make worker contributions.

Contributions for persons with no existing superannuation account are assumed to be paid into a "large" superannuation fund, defined as one with more than four members. Contributions are accumulated with investment earnings, after deducting expenses and any benefits paid. If a person is simulated as dying or emigrating, their account is assumed to be paid out in full, and closed. A person in a "large" fund can be simulated as "rolling into" a "small" fund (ie a fund with four or fewer members). Similarly, a person in a small fund can be simulated as "rolling out" into a large fund.

Persons aged 55 and over are assumed to be able to withdraw their whole balance and close the account. They can also choose to receive an allocated pension, which must be at least the statutory minimum proportion of their account balance for their age. No maximum now exists on allocated pension payments, but for these simulations it has been assumed that the total cash withdrawal in a year for a continuing account is 50% of the balance at the start of the year.

Allowance is made for annuities paid on an unfunded basis from large funds (generally public sector funds), and for a gradual fall in the numbers of new annuities. On the death of an annuitant who has a partner, a 60% reversionary pension is assumed to continue to the partner.

12.3 Administrative data sources 01-02 to 09-10

Published data used here were obtained as described below

- revenue accounts, total assets at the start an end of each year for 01-01 to 09-10 for all funds combined are from APRA (2011b)
- splits of total assets and member account numbers by fund type from 30/6/01 to 30/6/10 are from the same source
- splits by fund type of quarterly revenue account data for 01-02 to 03-04 are from APRA (2005)
- for funds with more than four members, splits by fund type of revenue account data from 04-05 to 09-10, and of the age distribution of vested benefits, were obtained from Superannuation Bulletins published annually by APRA (bulletins up to 07-08 are no longer available on APRA's website, and were obtained through a FOI request to APRA).

For 04-05 to 09-10, revenue account details for small funds were estimated by subtracting details for large funds from those for all funds. Pensions in 03-04 were shown in APRA 2011b as $13,127b, but have been here reduced to the $7.695b shown in APRA 2005 (thus removing an obvious anomaly).

From 07-08 on, the Australian Taxation Office (ATO) has collected annual details of the account balances, contributions, benefits and rollovers for each member of each self-managed fund. In response to a request under the Freedom of Information Act 1982, the ATO provided summaries by sex and 5-year age groups of these details for 07-08 and 08-09 (ATO 2011). As self-managed superannuation funds accounted for 99.3% of the total members of "small" funds at 30/6/10, the ATO data is a valuable guide to the behavior of small funds, as well as to rollovers into and out of small funds.

12.4 HILDA longitudinal data on superannuation contributions and assets

Questions about superannuation contributions and assets were included in HILDA waves 2 and 6. Unfortunately, questions about contributions were only asked about persons who described themselves as "not retired". Comparing HILDA-based estimates with APRA totals suggests that the former underestimated member contributions in waves 2 and 6 by about 50%. It seems likely that some recently-retired persons make large superannuation contributions, perhaps funded by the sale of their businesses. HILDA makes no distinction between self-managed and other funds, and its income receipt question groups together "superannuation/rollover fund/annuity/life insurance/allocated pension".

In spite of these severe limitations, regression models fitted to HILDA data are here used to impute base-line superannuation balances at 30/6/01, and to simulate employer and member contributions. Substantial adjustments to the contribution models are needed to align with APRA totals, particularly for years strongly affected by legislative changes.

12.5 Models of superannuation fund balances at 30/6/01

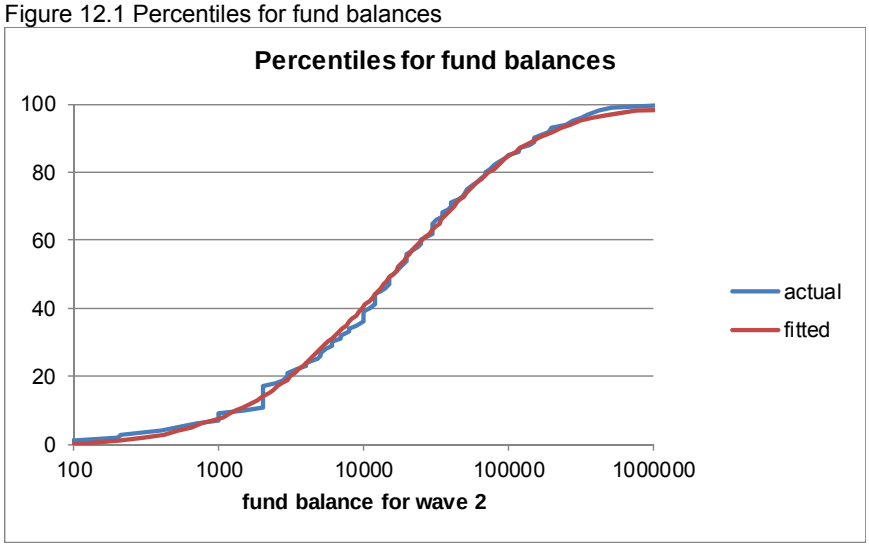
The models in table 12.1 were fitted to persons aged 15 and over in HILDA wave 2.

Table 12.1 Models of probability & size of fund balances

Variable	Probability			ln(Size)		
	Estimate	SE	Probability	Estimate	SE	Probability
age	0.147	0.007	0	0.164	0.007	0
agesquare	-0.00171	0.00008	0	-0.00109	0.000	0
earnings ($000)	0.0272	0.0016	0.0000	0.0126	0.001	0
occ1	1.191	0.122	0	0.238	0.067	0
occ2	1.383	0.095	0	0.275	0.049	0
occ3	1.465	0.104	0	0.280	0.055	0
occ4	1.251	0.105	0			
occ5	1.808	0.179	0	0.196	0.096	0.042
occ6	1.552	0.086	0			
occ7	1.240	0.117	0			
occ8	1.467	0.095	0	-0.274	0.067	0
occ9	1.016	0.094	0	-0.524	0.069	0
post	0.449	0.127	0	0.409	0.067	0
sex	-0.272	0.051	0	-0.436	0.035	0
tafe	0.305	0.057	0	0.153	0.037	0
under	0.473	0.062	0	0.196	0.040	0
constant	-3.181	0.158	0	5.097	0.148	0
N			13041			7904
LLR			0.312			
Adjusted R2						0.432
RMSE						1.438

The fitted values in figure 12.1 were obtained from the fitted linear estimates of the log(balance) by adding normally distributed random values with zero mean and standard deviation equal to the root mean square error, then exponentiating. The vertical step at $2,000 balance reflects about 7% of respondents reporting this as their approximate balance. In general, the simulated percentiles look reasonably close to the actual values.

Figure 12.1 Percentiles for fund balances



The logistic probability model in table 12.1 was used to randomly identify those persons in the baseline file having a non-zero fund balance. If simulated to have a balance, the size model in table 12.1 was used to randomly generate the amount of the log(balance), by adding normally distributed random values with zero mean and standard deviation equal to the root mean square error, then exponentiating. As the size model generated some very large fund balances, a maximum balance of $4m at 30/6/01 was assumed.

12.6 Probabilities of small fund membership at 30/6/01

The coefficients of the model in table 12.2 were chosen so as to balance with the known numbers and average sizes of small superannuation funds at 30/6/01, and to give an average age broadly consistent with the known ages of small fund members at 30/6/07.

Table 12.2 Logistic model of probability of fund at 30/6/01 being "small"

Variable	Assumed Coefficient
age	0.1
ln(fund)	0.3
constant	-10.88

12.7 Annuity recipients at 30/6/01

The 2000-01 Survey of Income and Housing Costs (ABS 2003b) asked for "current weekly income from superannuation/annuity/allocated pension". Table 12.3 shows the fitted logistic probability model for annuity receipt, together with a model for the probability of the annuity being only part of total income. For those cases where the annuity was less than total income, the last two columns shows a linear regression model fitted to log(annuity/(income - annuity).

Table 12.3 Models for annuity receipt and size of annuity at 30/6/01

Coefficient	Annuity receipt Estimate	Annuity receipt SE	Partial annuity Estimate	Partial annuity SE	Annuity proportion Estimate	Annuity proportion SE
age	0.685	0.054			0.270	0.076
agesquare	-0.00472	0.00041	0.00041	0.00014	-0.00215	0.00059
incomepw	0.00046	0.00007	0.00263	0.00082	-0.00042	0.00017
sex	-0.683	0.095	0.780	0.374	-0.382	0.161
constant	-25.401	1.769	-1.396	0.869	-7.078	2.485
N		13183		573		530
LLR		0.269		0.079		
Adjusted R2						0.040
RMSE						1.745
Adjustment	-0.425					

The weighted estimate of Australian pensions and annuities obtained from the SIHC was $7.83b. By contrast. superannuation pensions in 00-01 were reported as $6.44b (APRA 2011b). The difference may reflect the payment of some annuities outside the superannuation system, as well as sampling error. The adjustment shown above to the constant for the model for the probability of receiving an annuity was made to give superannuation annuities plus allocated pensions in 01-02 closely balancing with the $6.849b reported by APRA for 01-02.

12.8 Investment earning and expense rates

Table 12.4 shows earning rates estimated from APRA data as

(net investment earnings) / (total assets at start + .5 * (net cash inflow in year)

where net cash inflow is contributions and rollins, less benefits, rollouts and expenses.

Also shown are adjusted earning rates, where the net investment earnings have been adjusted to eliminate any misbalances between opening and closing total funds. For all but one year, the estimated earning rates are much higher for small funds than large, with an average difference over the 9 years of 6.6% pa. This large difference is surprising, as small funds are likely to have higher expense rates and more conservative investments. Table 12.4 also shows expense rates, estimated as the expenses In a year divided by the mean of total assets at the start and end of the year.

Table 12.4 Investment earning and expenses rates

Year ending 30 June	Earning rate Large	Earning rate Small	Adjusted Earning Rate Large	Adjusted earning rate Small	Expense rate Large	Expense Rate Small
2002	-4.2%	11.7%	-7.6%	4.9%	0.60%	1.52%
2003	-2.2%	13.2%	-1.1%	5.0%	0.59%	1.53%
2004	15.1%	12.2%	8.9%	9.9%	0.57%	1.42%
2005	11.8%	17.9%	11.1%	16.4%	0.49%	1.68%
2006	12.9%	16.6%	12.1%	16.4%	0.48%	1.65%
2007	14.4%	21.8%	13.4%	21.6%	0.47%	1.75%
2008	-7.3%	-6.4%	-9.6%	-4.4%	0.47%	1.13%
2009	-10.7%	-3.0%	-12.3%	-1.8%	0.49%	1.06%
2010	8.3%	13.1%	8.3%	12.1%	0.52%	1.02%
Average	4.2%	10.8%	2.6%	8.9%	0.52%	1.42%
Future	5.5%	5.5%			0.52%	1.02%

12.9 Employer contribution rates greater than 9%

The logistic probability model in table 12.5 was fitted for the 9,839 employees in HILDA waves 2 and 6 with known employer contribution rates. Of these, 1,254 reported employer contribution rates greater than 9%, the statutory contribution rate that has been required since 1/7/02. The probability model was used to simulate the probability of employer contributions greater than 9%, and the logarithmic model was used to simulate the size of the excess contribution over 9%.

Table 12.5 Models of probability and size of employer contribution rates >9%

Variable	Probability			ln(rate-9%)		
	Estimate	SE	Probability	Estimate	SE	Probability
age				0.009	0.002	0
agesquare	0.00011	0.00003	0			
occ1	0.653	0.116	0	0.256	0.085	0.003
occ2	0.638	0.089	0	0.364	0.065	0
occ3	0.578	0.102	0	0.183	0.077	0.017
occ4				0.315	0.096	0.001
occ5	0.534	0.190	0.005			
occ6	0.286	0.109	0.009			
sex	-0.480	0.070	0			
wagereal $000	0.007	0.001	0			
wave6	0.221	0.062	0			
constant	-2.263	0.130	0	0.389	0.095	0
N			9839			1254
LLR			0.041			
Adjusted R2						0.046
RMSE						0.885

12.10 Employer contribution rates less than 9%

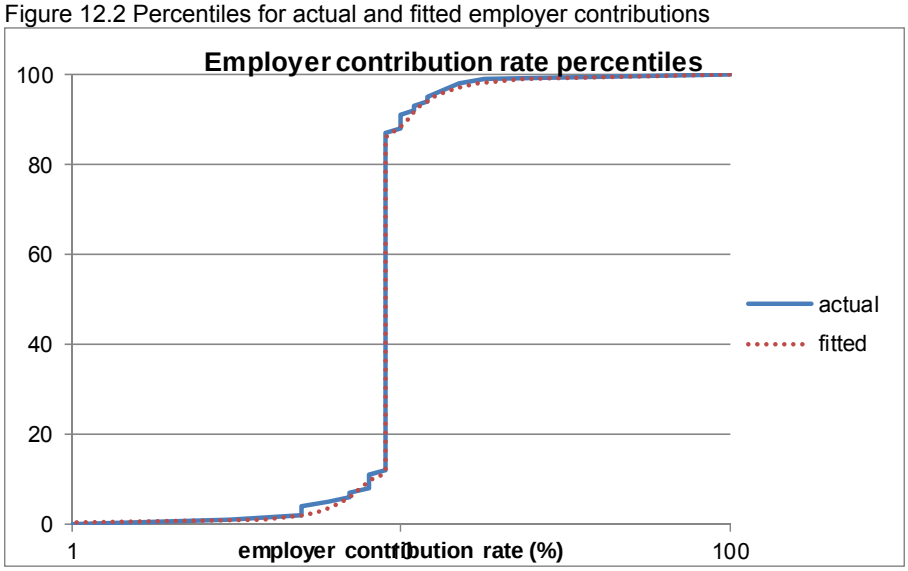
The logistic probability model in table 12.6 was fitted for the 8,585 employees with employer contribution rates less than or equal 9%. Of these, 1,123 reported employer contribution rates less than 9%.

Table 12.6 Models of probability and size of employer contribution rates <9%

Variable	Probability			ln(9%-rate)		
	Estimate	SE	Probability	Estimate	SE	Probability
age	0.039	0.017	0.022	-0.022	0.011	0.039
agesquare	-0.00056	0.00022	0.009	0.00033	0.00014	0.018
constant	-2.465	0.314	0	1.177	0.201	0
N			8585			1123
LLR			0.0017			
Adjusted R2						0.007
RMSE						0.687

The fitted values in figure 12.2 were obtained from the size models in tables 12.5 and 12.6, by adding normally distributed random values with zero mean and standard deviation equal to the root mean square error, then exponentiating. As shown by the HILDA data, 76% were assumed to have the statutory contribution rate of 9%.

Figure 12.2 Percentiles for actual and fitted employer contributions



12.11 Models of probability and size of worker contributions

The logistic probability model in table 12.7 was fitted for the 16,364 employed persons in HILDA waves 2 and 6. Of these, 2,799 had recorded worker contributions. In applying the size model, cases with simulated negative and zero contributions were assumed to not make a worker contribution.

Table 12.7 Models of probability and size of worker contributions

Variable	Probability			Size		
	Estimate	SE	Probability	Estimate	SE	Probability
age	0.240	0.014	0	-656	197	0.001
agesquare	-0.0023	0.0002	0	9.7	2.3	0
busincreal $000	0.012	0.002	0	43	12	0
occ1				-5013	2501	0.045
occ2				-6850	2420	0.005
occ3	-0.155	0.067	0.022	-5094	2476	0.04
occ4	-0.285	0.077	0	-7460	2497	0.003
occ5				-8261	2836	0.004
occ6	-0.177	0.068	0.01	-8043	2462	0.001
occ7	-0.607	0.095	0	-7796	2616	0.003
occ8	-0.680	0.122	0	-9070	2827	0.001
occ9	-0.819	0.113	0	-8813	2738	0.001
sex	-0.395	0.051	0			
wagereal $1000	0.011	0.001	0	45	9	0
wave6	-0.240	0.045	0			
constant	-6.749	0.302	0	18486	4666	0
N			16364			2799
LLR			0.1193			
Adjusted R2						0.0381
RMSE						15053

12.12 Adjustments to contribution models

The adjustments in table 12.8 were derived by repeated trials, finding adjustments which reasonably reproduced the actual contributions in each year (see figures 12.3

and 12.4). Years 1 to 5, from 01-02 to 05-06, were grouped, as there were no major changes to relevant legislation. Years 8 and 9, 08-09 and 09-10, were similarly grouped. The adjustments were added to the constant terms of the models of the probability of the contribution occurring, and also added to the constant terms of the models of log(rate - 9%) for the models of employer contributions greater than 9%. For the models of member contributions, the adjustment was exponentiated, and used to multiply the simulated contributions. The adjustments were iteratively calculated, using the formula

New adjustment = Adjustment used in simulation + ln(ratio of actual to simulated contributions)

Table 12.8 Adjustments to constant terms of contribution models

Model	Years	1-5	6	7	8-9
Employer contributions >9%	Large	0.35	1.11	1.04	1.00
Employer contributions >9%	Small	3.01	3.63	3.24	2.78
Member contributions	Large	-0.28	0.09	-0.24	-0.74
Member contributions	Small	1.37	2.60	2.19	1.64

Large adjustments were needed to balance with employer and member contributions to small funds. This may reflect the incompleteness of the HILDA contribution data from which the models have been derived. The largest adjustments were needed in year 6 (06-07), when major changes had been foreshadowed, and when unusually large contributions were temporarily permitted. No explicit allowance was made in the simulations for the 15% taxes deducted from employer contributions, but the adjustments ensured a reasonable balance with the after-tax APRA data.

12.13 Actual and simulated contributions 1/7/01-30/6/10

To reduce the effects of random variations, the simulated values in figures 12.3 and 12.4 are the averages of 5 runs. Such averaging is particularly necessary for low-probability events with skew size distributions, such as member contributions.

Figure 12.3 Large fund contributions

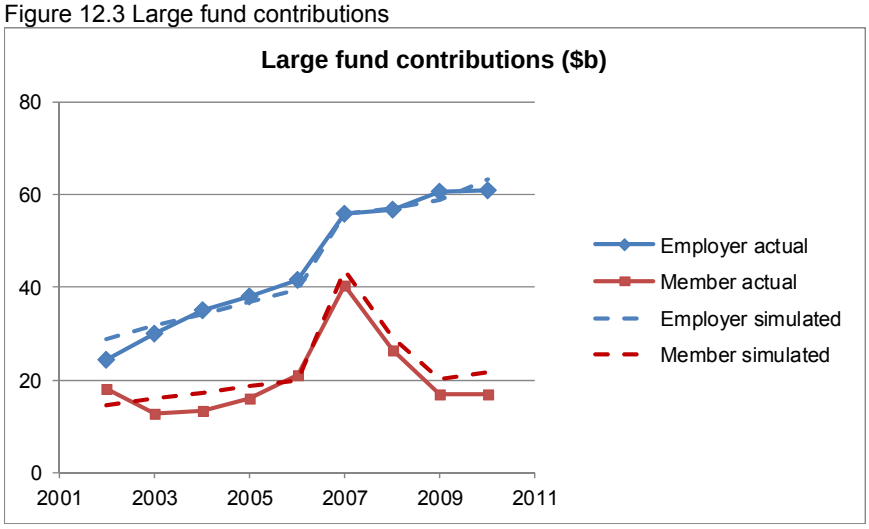
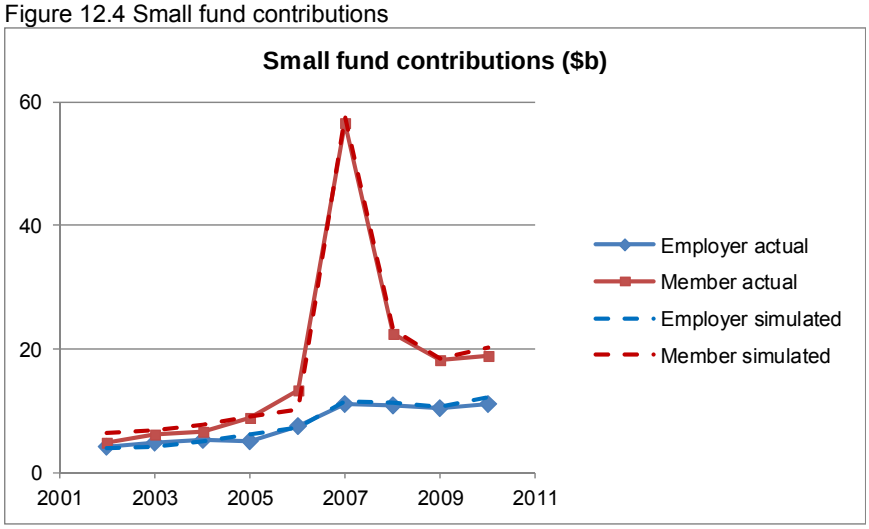


Figure 12.4 Small fund contributions



Figures 12.3 and 12.4 show the dramatic increases in member contributions in 06-07, approximately doubling in large funds and tripling in small funds. In 06-07 the ending of most limits and taxes on superannuation benefits had been announced, and the non-concessional contribution cap was due to be reduced from $1m to $0.15m on 30/6/07.

The deviations between actual and simulated contributions in the first five projection years are hard to explain, but are not of consequence for projections after 09-10. As contribution levels appear to have stabilized for employer and member contributions, for both self-managed and other funds, it may be reasonable to assume that the model and adjustments for the 09-10 year will be useful for projecting future years.

12.14 Account closure assumptions

Lump sum usage rates were approximately estimated from the detailed statistics for self-managed super funds in 07-08 and 08-09 (ATO 2011). These rates are intended as estimates of the probabilities of member balances ceasing to exist, as a result of death, emigration or account closure. Estimated mortality rates were deducted from the averages of the lump sum usage rates. Emigration rates are low at older ages, so no adjustment was made for them here. The resulting rates were assumed as account closure rates for ages 55-69, for small superannuation funds in all projection years. To balance approximately with the observed lump sum payments from large superannuation funds, they were assumed to have twice the closure rates. No closures were assumed below age 55, or above 69. Assumed closure rates are in table 12.9.

Table 12.9 Closure assumptions derived from SMSF experience

Age	Lump sum usage 07-08	Lump sum usage 08-09	Lump sum usage Average	Mortality rate	Lump sum usage - Mortality	Assumed closure rates
55-	0.0116	0.0102	0.0109	0.0032	0.0077	0.0077
60-	0.0438	0.0363	0.0401	0.0059	0.0342	0.0342
65-	0.0315	0.0170	0.0242	0.0108	0.0134	0.0134
70-	0.0085	0.0127	0.0106	0.0261	-0.0155	0.0000

12.15 Allocated pension assumptions

Persons were assumed to be able to convert to allocated pensions from age 55 on, with probabilities given by the models in table 12.10. The constant in each year was chosen so as to reasonably reproduce actual pension benefits.

Table 12.10 Logistic model for allocated pension entry

Type	Year	1-4	5	6	7	8-9
Large	Age	0.01	0.01	0.01	0.01	0.01
Large	Constant	-6.50	-2.00	-4.00	-2.00	-2.00
Small	Age	0.01	0.01	0.01	0.01	0.01
Small	Constant	-6.50	-1.80	-4.00	-1.20	-1.20

Up to 30/6/07, legal minimums and maximums, specified as percentages of the fund balance at the start of the financial year, applied to allocated pension withdrawals. As from 1/7/07, lower minimum percentages applied, and no maximums. Applicable minimums and maximums were assumed during the simulations, with a maximum of 0.5 if no legal maximum applied (ATO 2011b).

Excesses over the legal minimum pension rate were calculated as -0 .04 * ln r, where r is a N(0,1) random variable. For 07-08 to 09-08, where the minimum pension percentages were temporarily halved, the 0.04 average excess was increased to 0.05, and for 11-12, where the minimums were 75% of the long-term values, the 0.04 excess was increased to 0.045.

No data were available on the distribution of allocated pension withdrawals, but the assumed mean excesses are broadly consistent with the pension numbers and magnitudes shown by the detailed statistics for self-managed super funds in 07-08 and 08-09 (ATO 2011a).

12.16 Annuity assumptions

Existing annuities were assumed to increase in line with average weekly earnings each year. Persons without annuities aged 55 to 64 were assumed to have a 0.01 probability of starting an annuity in 01-02, with this probability reducing by 10% each subsequent projection year. The sizes of new annuities were assumed to be log-normally distributed, with mean and standard deviation based on the 00-01 data for persons receiving weekly income from superannuation/annuity/allocated pension, allowing for subsequent average weekly yearnings inflation. On the death of persons with a partner, a 60% reversion was assumed to the partner. The validity of all these assumptions is uncertain, and more data are needed to improve them.

12.17 Benefit projections

Simulated lump sums allow for the full payment of the fund balance on death or emigration, as well as on voluntary closure. Figure 12.5 shows simulated benefits for large funds, and figure 12.6 shows benefits for small funds. The assumed closure rates for large funds are twice the rates derived from SMSF experience in 07-08 and 08-09. The peak in 07-08 actual lump sums may reflect withdrawals to make additional contributions. It is not clear why actual lump sums in 01-02 and 02-03 are so much higher than the simulated values. Simulated pensions include annuities, and are reasonably close to actual.

Figure 12.5 Large fund benefits

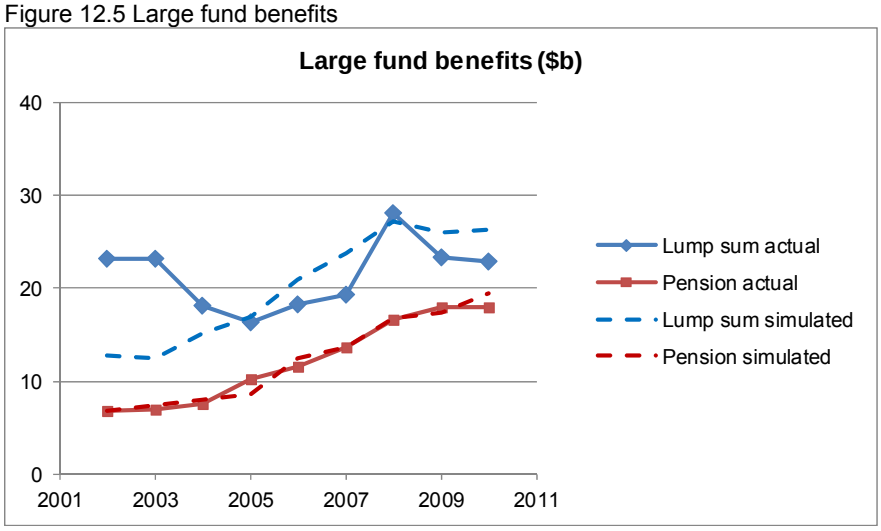
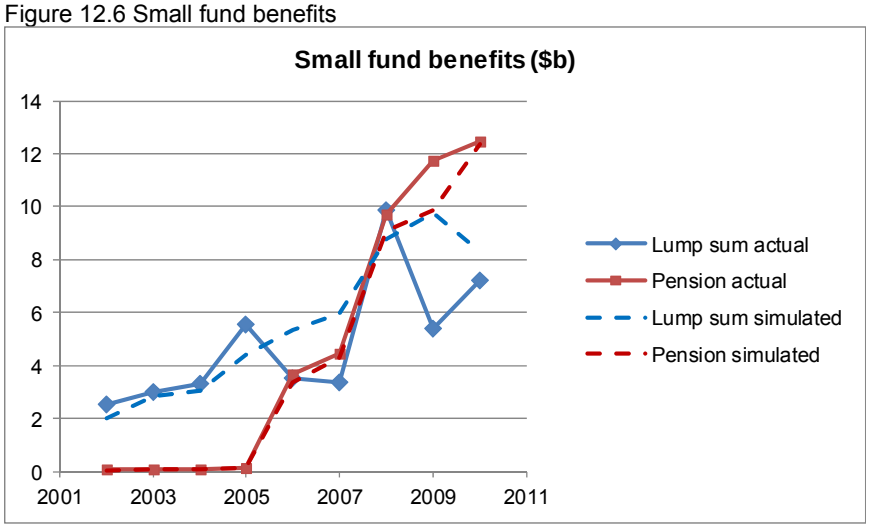


Figure 12.6 Small fund benefits

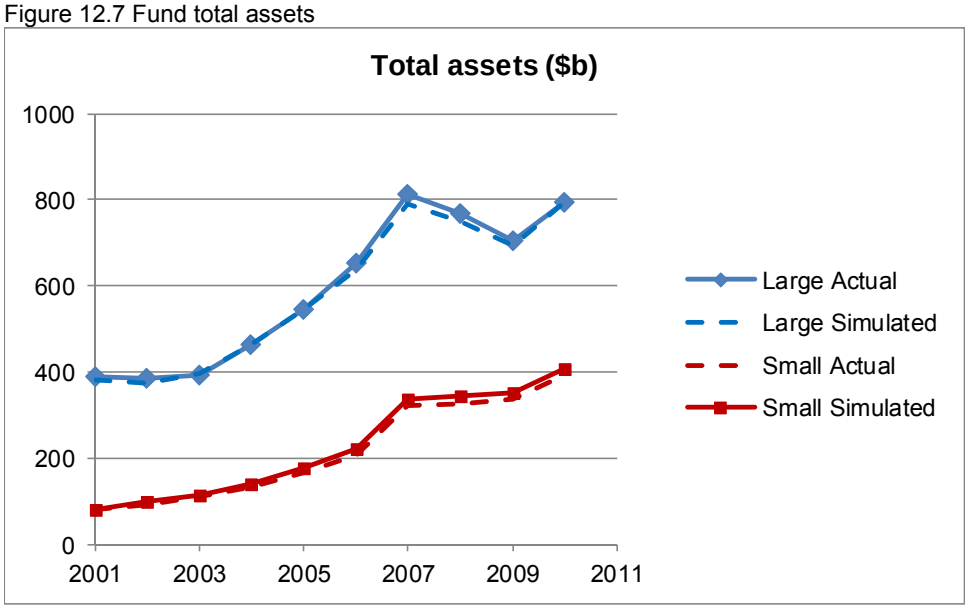


Simulated lump sums for small funds are broadly based on SMSF experience in 07-08 and 08-09. As for large funds, the peak in 07-08 actual lump sums may reflect withdrawals to make additional contributions. The simulated pensions are based on the entry assumptions in table 12.10, intended to reasonably replicate actual pension payments. No annuities are included in the simulated pensions for small funds.

12.18 Fund total assets and average ages

Assumption adjustments were chosen so as to give initial assets, contributions, benefits, rollovers, investment earning rates and expenses broadly matching actual figures. The closeness between actual and simulated assets in figure 12.7 is the result of this matching.

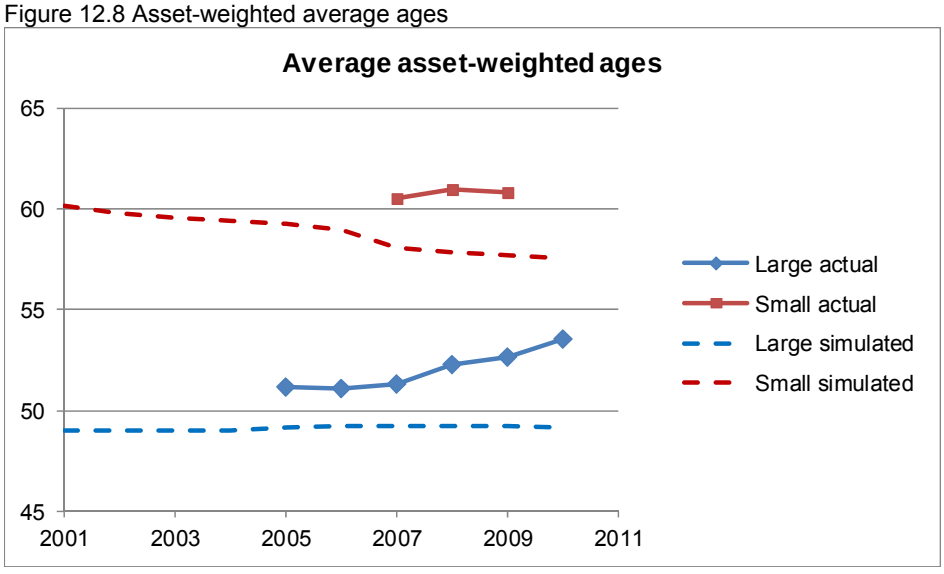
Figure 12.7 Fund total assets



12.19 Member average ages

APRA Superannuation Bulletin data for large funds gives vested benefits subdivided into ages <35, 35-49, 50-59, 60-65 and >65. The average asset-weighted ages for large funds in figure 12.8 are thus approximate estimates. Ages for small funds are from 5-year age groups (ATO 2011) and thus more reliable.

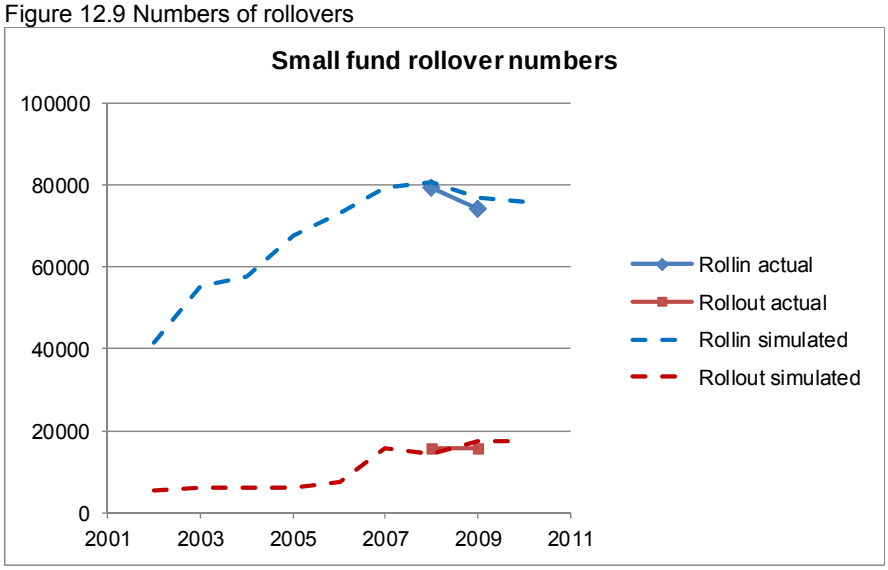
Figure 12.8 Asset-weighted average ages



12.20 Numbers of rollovers into and out of small funds

Assumptions about rollovers into and out of small funds were chosen to match the two years of data available (ATO 2011), and to approximately match total member account numbers. Figure 2.9 shows that good matches were achieved for the only two years for which data were available. No comparisons with large fund rollover numbers could be made, as these include many rollovers from one large fund to another.

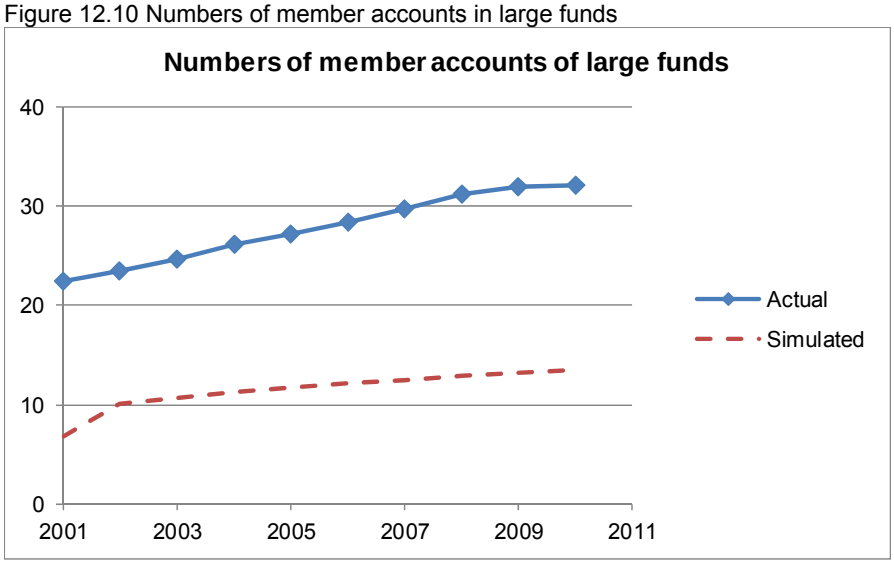
Figure 12.9 Numbers of rollovers



12.21 Numbers of member accounts in large funds

Large funds are known to contain many duplicate member accounts. The simulations assume that a person with superannuation has only one account, either in a large or small fund. The simulated numbers of accounts in large funds were about 31% of the published numbers at 30/6/11, rising to 43% a year later, and staying at about that level for the rest of the simulation (see figure 12.10). This suggests that the HILDA-based model used to impute numbers of persons with superannuation accounts at 30/6/01 underestimated the actual numbers of members by about 28%. In the simulations, anyone who receives an employer contribution, or makes a worker contribution, is assumed to open an account in a large fund if they do not already have an account.

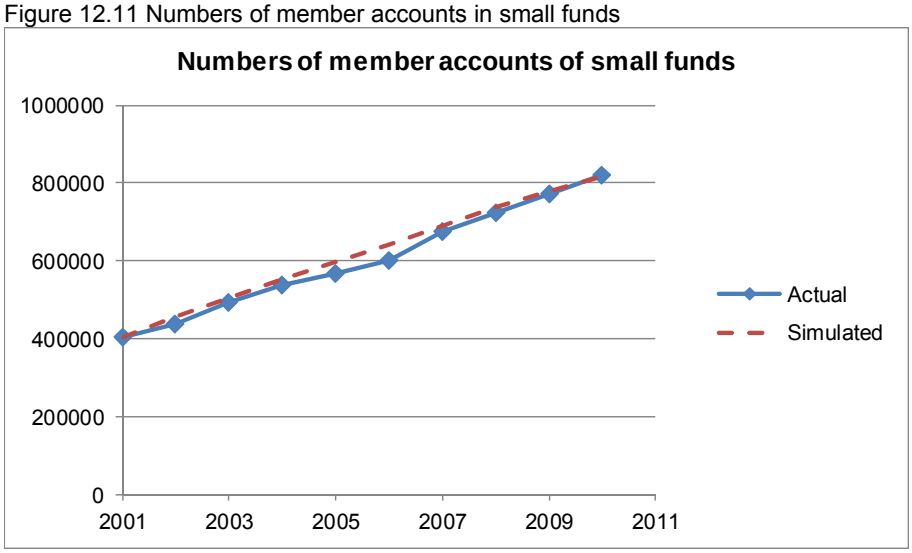
Figure 12.10 Numbers of member accounts in large funds



12.22 Numbers of member accounts in small funds

The reasonable fit between actual and simulated account numbers in figure 12.11 is the result of adjustments to initial imputations, rollover rates and account closure rates.

Figure 12.11 Numbers of member accounts in small funds



12.23 Conclusions

HILDA data on superannuation were only available for 02-03 and 06-07, contained no benefit information, and seriously underestimated contributions by omitting the retired. Models derived from the HILDA data proved useful in imputing initial account balances, and in modelling employer and worker contributions. Substantial adjustments to the contribution models were needed to match APRA contribution data, which showed the large effects of major legislative changes. An iterative process was used to find these adjustments.

Data on account balances, contributions, benefits and rollovers were available for each member of a self-managed fund in 07-08 and 08-09 (ATO 2011). Using these data, and allowing for relevant legislation, models of lump sum benefits, allocated pensions and rollovers were estimated, and adjusted to balance with the APRA data for each year.

Fund earning and expense rate assumptions were based on APRA data. Surprisingly, in the 9 years to 30/6/10, small funds averaged investment returns of 10.8% pa, compared with 4.2% for large funds.

Although based on a patchwork of survey and administrative data, the modelling does appear to be doing a reasonable job of fitting the experience in the 9 years to 30/6/10. Given the major legislative changes in the last 5 years, it may take several more years for the superannuation system to stabilize, and for reliable projections to be feasible.

As superannuation is now the second-largest type of private investment, it is important that any future legislative changes be based on reasonable projections.

13. HOUSING

Unusually, the Cumpston model treats dwellings and households separately. Together with the 57 region structure, this allows detailed simulations of the dwellings chosen by moving households.

The 2001 census provided tenure data, together with repayments within ranges for dwellings being bought, and rentals within ranges for rented dwellings. These range numbers had to be converted to continuous values, and missing values filled. Dwelling market values were not available, and had to be imputed using survey data. Some census overstatement of ownership was evident amongst young heads of households, and had to be approximately corrected.

Simulations of moves by households, and by persons wanting to form households, relied on six key models

- destination region
- tenure type in the new household (three models)
- value of a dwelling occupied by a household
- loan to value ratios for occupied dwellings with mortgages.

As the available tenure data were for occupied dwellings rather than moving households, an iterative process was used to adjust the tenure models so as to approximately replicate 09-10 survey data on tenure by age groups. With one exception, this iterative process worked well.

To ensure reasonable vacancy levels, dwelling building and demolition rates were assumed for each region.

13.1 Structure of Cumpston model

Each private dwelling, whether occupied or not, is assumed to have a market value. Each occupied private dwelling is assumed to have one of four tenure types - fully owned, owned with a mortgage, rented, or occupied rent-free.

A tenure type is selected for moving households, and for persons wanting to establish new households. A destination region is selected, taking into account the source region and the age of the head of the household. If there are at least three vacant dwellings, the most suitable dwelling is selected, taking into account the characteristics of the dwelling and household. If there are less than three vacant dwellings, another destination region is selected.

If the dwelling is to be rented, a rental rate is calculated based on regional rental yields. If the dwelling is to be owned with a mortgage, the term of the mortgage and the initial loan to value ratio are chosen based on the age of the head of household, and the monthly repayment calculated based on current variable loan rates.

The outstanding balance on each mortgage is periodically recalculated, and the tenure changed to fully owned if the mortgage has been repaid. Rents are recalculated, taking into account current regional rental yields.

Depending on the vacancy rate in each of the 57 regions, developers are assumed to build some new dwellings, and other dwellings are assumed to be demolished. These building and demolition processes are intended to provide sufficient local vacancies.

13.2 Assumed gross rental yields

The gross rental yields in table 13.1 were obtained from net rental yields (Real Estate Institute of Australia 2008), by dividing by the 0.8 assumed by the Institute to estimate net yields. The June 2001 yields were used to estimate dwelling values from rents recorded in the 2000-01 SIHC, and to estimate dwelling values from rents recorded in the HSF. Rental yields from June 2002 on were used to estimate rents for dwellings simulated as rented.

Table 13.1 Assumed gross rental yields (percentages)

June	Sydney	Melbourne	Brisbane	Adelaide	Perth	Hobart	Darwin	Canberra
2001	2.69	2.79	4.94	5.19	4.07	5.53	5.34	4.71
2002	2.21	2.79	4.72	4.53	3.85	5.60	4.99	4.57
2003	2.00	2.43	3.90	3.59	3.50	4.28	5.01	3.82
2004	1.88	2.56	3.04	3.33	3.22	3.47	4.08	3.34
2005	2.05	2.64	3.22	3.39	3.08	3.52	4.24	3.54
2006	2.11	2.55	3.57	3.48	2.63	3.83	3.57	3.50
2007-	2.20	2.50	3.30	3.40	2.80	3.60	4.20	3.40

13.3 Regression model for ln(value) of occupied dwellings in 2001

The logistic model in table 13.2 was derived from the 6,624 2000-01 SIHC households for which sale value was known or derived from rent. “Income” is the income of the household, as a proportion of the average household income. “Age” is the age of the household reference person. State price differentials are relative to NSW with “other” being the Northern Territory and the Australian Capital Territory.

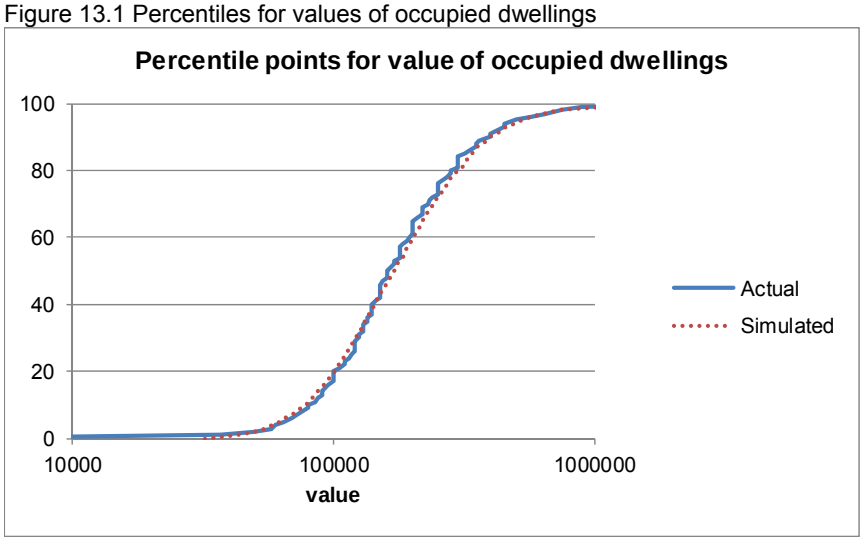
Table 13.2 Regression model for ln(value) of occupied dwellings

Variable	Estimate	SE	P>|z|
Age	-0.00169	0.00051	0.001
Buying	0.125	0.019	0
Capital	0.366	0.015	0
Income	0.187	0.008	0
Numperson	0.046	0.006	0
Owned	0.259	0.020	0
Separate	0.102	0.019	0
Vic	-0.298	0.020	0
Qld	-0.520	0.021	0
SA	-0.712	0.023	0
WA	-0.436	0.023	0
Tas	-0.804	0.029	0
Other	-0.174	0.032	0
Constant	11.680	0.034	0
N			6624
Adjusted R-squared			0.371
Rmse			0.534

Fully owned dwelling were more valuable than dwellings being bought, and both were more valuable than rented dwellings. Higher income households, and those with more persons, were found to occupy more valuable dwellings. Dwellings in capital cities were more valuable than those outside capitals.

The simulated values in figure 13.1 were obtained from the fitted linear estimates of the log(value), by adding normally distributed random values with zero mean and standard deviation equal to the mean of the regression root mean square error.

Figure 13.1 Percentiles for values of occupied dwellings



13.4 Estimation of rents, dwelling values, loans and repayments at 30/6/01

For rented dwellings, the 1% census sample provided weekly rents in 16 size bands. For the bottom 15 bands, rents were randomly selected within the band. For the top band, for rents of $500 per week and over, rents were randomly selected from an assumed size distribution with average $593. If the rent was not stated, one of the 16 size bands was randomly selected.

For dwellings being bought, the census sample provided monthly loan repayments in 15 size bands. For the bottom 14 bands, repayments were randomly selected within the band. For the top band, for repayments of $1500 per month and over, rents were randomly selected from an assumed size distribution with average $1778. If the repayment was not stated, one of the 15 size bands was randomly selected.

For dwellings owned, being bought or occupied rent-free, the market value was imputed from the regression model in table 13.2. Estimates of the log(value) were calculated from the regression relationship, adding normally distributed random values with zero mean and standard deviation equal to the mean of the regression root mean square error.

For rented dwellings, market value was estimated from weekly rent, using the gross rental yields for June 2001 in table 13.1.

13.5 Adjustments to allow for overestimation of ownership at 30/6/01

Comparing tenure proportions in the baseline file for each age group with those reported by the 2000-01 SIHC, the proportions reporting ownership or buying at young ages appeared too high. The reported numbers of owners aged up to 44 were thus reduced by the proportions in table 13.3, with the dwellings assumed to be being bought rather than owned outright. One-third of buyers in the 15-24 age group were assumed to be renters. For dwellings changed from owned to being bought, mortgage terms, loans and repayments were estimated as for moving households. For dwellings

changed from being bought to rented, rentals were estimated as for moving households.

Table 13.3 Proportion of owned houses assumed to be being bought

Age of head	15-24-	25-34	35-
Proportion of owners assumed to be buyers	0.82	0.38	0.22
Proportion of buyers assumed to be renters	0.33		

If there are less than three vacant dwellings in the destination region, another destination region is selected. The use of three dwellings was indicated by early trials (Cumpston 2009a), but requires further investigation.

13.6 Loan to value ratios for dwellings being bought

There were 2,165 2000-01 SIHC households for which loans were being repaid. Of these, 42 had loan to value ratios of 1.2 or higher, and these were assumed to be errors. In practice, lenders would be unwilling to tolerate ratios much higher than 1, and would foreclose. The model in table 13.5 was obtained from the remaining 2,123 cases, where the regressed variable was

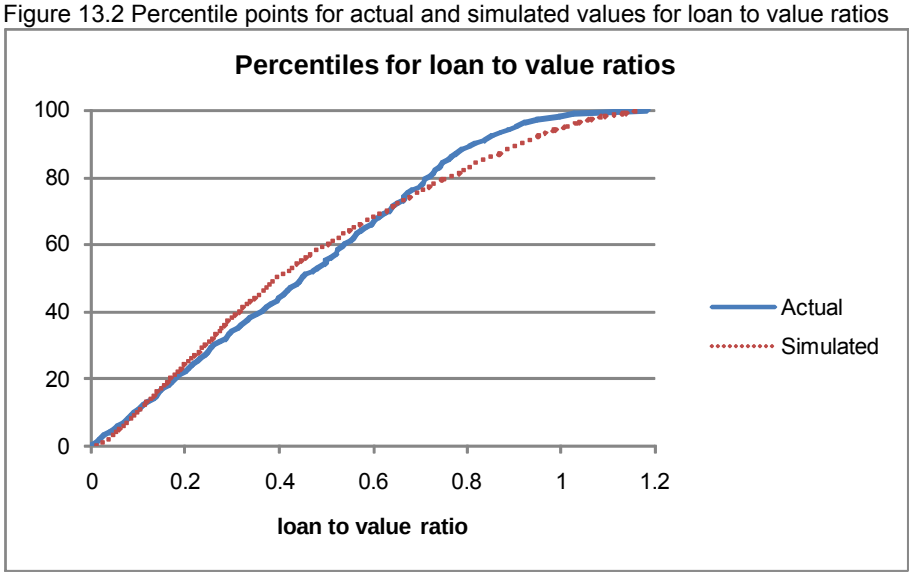
$$y = \ln(r / (1 - r)) \qquad \text{where } r = (\text{loan to value ratio}) / 1.2$$

Table 13.5 Model of transformed loan to value ratio for dwellings being bought

Variable	Estimate	SE	P>|z|
age	-0.0294	0.0029	0
years	-0.16608	0.01753	0
yearssquare	0.00678	0.00122	0
yearscube	-0.00011	0.00002	0
capital	-0.225	0.053	0
constant	1.424	0.126	0
Adjusted R-squared	0.261		
Root MSE	1.190		

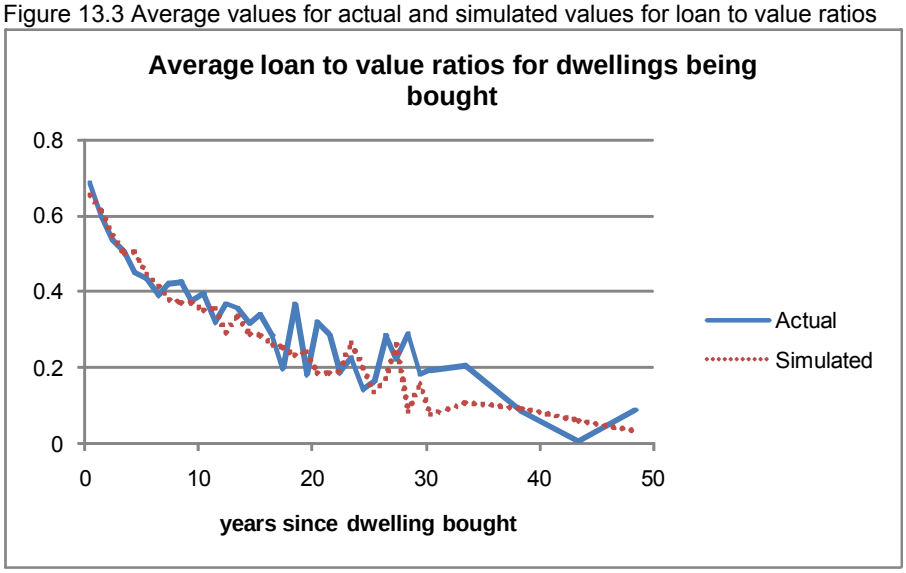
The simulated values in figure 13.2 were obtained from the fitted linear estimates, by adding normally distributed random values with zero mean and standard deviation equal to the root mean square error of the regression, and reversing the transformation. The fit is mediocre, perhaps reflecting some errors and the arbitrary nature of the assumed 1.2 limit.

Figure 13.2 Percentile points for actual and simulated values for loan to value ratios



After a dwelling has been bought with a loan, the outstanding loan balance should slowly decrease, and the value of the dwelling should increase. The continuing drops in loan to value ratios in figure 13.3 thus seem reasonable, and the actual and simulated values are reasonably close. Fluctuations from about 15 years after purchase probably reflect low numbers of dwellings.

Figure 13.3 Average values for actual and simulated values for loan to value ratios



13.7 Annual updating of dwelling values, mortgage balances, loan repayments and rents

The estimated value of each private dwelling is updated each June, using the assumed dwelling value increases in table 13.6.

Table 13.6 Assumed dwelling value increases (percentages)

June	Sydney	Melbourne	Brisbane	Adelaide	Perth	Hobart	Darwin	Canberra
2002	21.7	19.0	21.3	17.0	10.8	7.3	3.2	16.7
2003	14.8	12.6	28.8	23.5	16.8	38.4	11.9	28.6
2004	4.3	4.7	24.9	13.1	16.3	35.3	12.8	10.3
2005	-3.6	4.0	1.6	5.3	16.8	6.2	19.0	-0.3
2006	0.1	6.4	5.1	5.6	38.4	8.8	22.6	7.0
2007	4.1	13.7	15.5	11.5	13.3	8.7	10.6	10.7
2008	3.0	14.5	14.1	15.8	-0.7	5.7	6.9	6.9
2009	-0.8	0.8	-2.7	1.4	-2.9	1.3	11.1	-0.2
2010	16.9	22.8	8.5	9.3	12.4	7.7	13.2	16.0
2011	-0.7	-2.0	-3.6	-2.1	-4.1	2.8	-3.0	2.2
2012-	2.8	7.8	7.5	7.3	10.4	9.1	11.4	6.4

The value increases up to June 2011 are for established houses in capital cities (ABS 2011b). The assumed increases for June 2012 and on are the averages of the years ending June 2004 to 2011 (thus leaving out the high growth years to June 2002 and 2003). Rents are recalculated from the new values, taking into account current regional rental yields (table 13.1).

The outstanding balance on each mortgage is recalculated at the end of each projection year, based on the balance at the start of the year, the housing loan interest rate (table 13.7) and the monthly repayment. The tenure is changed to fully owned if the mortgage has been repaid.

The variable loan rates in table 13.7 are standard bank variable interest rates for housing loans in June (RBA 2011).

Table 13.7 Assumed variable loan rates (percentages)

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Rate	6.55	6.55	6.55	7.05	7.30	7.55	8.05	9.45	5.80	7.40	7.80

13.8 Selecting dwellings for moving households

Using the logistic models in tables 13.7 to 13.9, a tenure type is selected for moving households, and for persons wanting to establish new households. Using the logistic regression models in table 9.5, a destination region is selected, taking into account the source region and the age of the head of the household.

If there are at least three vacant dwellings in the destination region, the fitted value of each of three randomly selected dwellings is calculated using the regression model in table 13.2. Of the three dwellings evaluated, the one whose ln(fitted value) is closest to its ln(actual value) is chosen as the new dwelling for the household.

If the tenure type is simulated as renting, the weekly rental is estimated from the current value of the dwelling and current rental yields for dwellings (table 13.1).

If the dwelling is to be bought with a mortgage, the mortgage term is assumed to be age 65 less the age of head of household, with a minimum term of 5 years and a

maximum term of 30 years. Using the regression model in table 13.5, the initial loan to value ratio is simulated, allowing for random normal distribution of the transformed ratio with standard deviation equal to the root mean square error of the regression. The initial mortgage is calculated from the dwelling value and the loan to value ratio, and the monthly mortgage repayment calculated assuming the current housing variable interest rate (table 13.7).

13.9 Probability of rent-free housing for moving households

The logistic model in table 13.7 was derived from all 6,786 households in the 2000-01 SIHC. Of these, 130 were not renting, buying or owning, and were treated here as rent-free. Age was taken as the age of the reference person in the household. "Income" was calculated as household income divided by average household income. To balance approximately with the tenure proportions for each age group estimated for 09-10 by ABS (2011a table 14A), the adjustments in table 13.10 were made to the constant term of the model.

Table 13.7 Logistic model of probability of not renting, buying or owning

Variable	Estimate	SE	P>|z|
Age	-0.072	0.026	0.006
agesquare	0.00057	0.00025	0.025
Income	-1.088	0.175	0
loneparent	-1.052	0.433	0.015
loneperson	0.388	0.180	0.031
Constant	-0.995	0.636	0.118
N			66786
LLR			0.061

13.10 Probability of renting if not rent-free for moving households

The logistic model in table 13.8 was derived from the 6,624 2000-01 SIHC households for which sale value was known or derived from rent. Household income, here measured as a proportion of average household income, has a strong negative influence on the probability of renting. Probabilities of renting also decrease strongly with the age of the reference person in the household. To balance approximately with the tenure proportions for each age group estimated for 09-10 by ABS (2011a table 14A), the adjustments in table 13.10 were made to the constant term of the model.

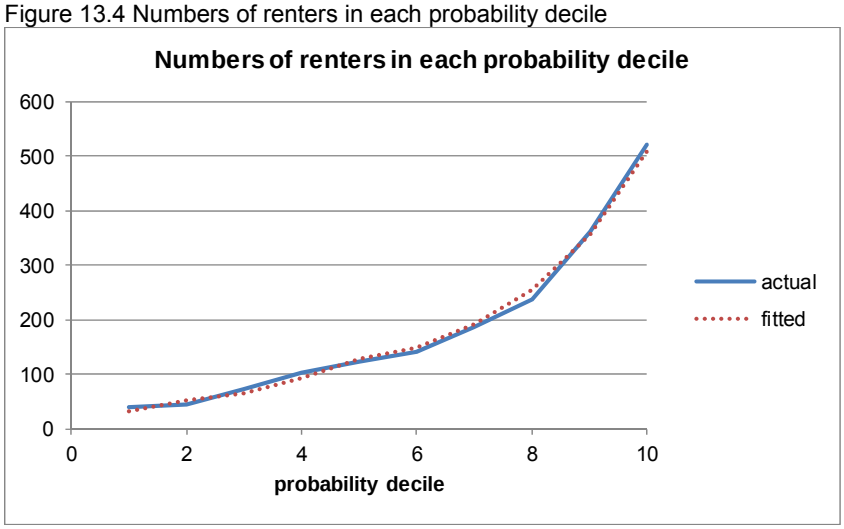
Table 13.8 Logistic model of probability of renting

Variable	Estimate	SE	P>|z|
age	-0.153	0.012	0
agesquare	0.00095	0.00012	0
group	1.344	0.159	0
income	-0.554	0.058	0
loneparent	1.315	0.106	0
loneperson	1.054	0.080	0
constant	3.868	0.293	0
N			6624
LLR			0.203

The 6,624 dwellings were sorted in order of their fitted probability of being rented, and grouped into deciles. Figure 13.4 shows the actual number of rented dwellings in each

decile, together with the sum of the fitted probabilities for the decile. The fit appears to be good.

Figure 13.4 Numbers of renters in each probability decile



13.11 Probability of buying if buying or owning for moving households

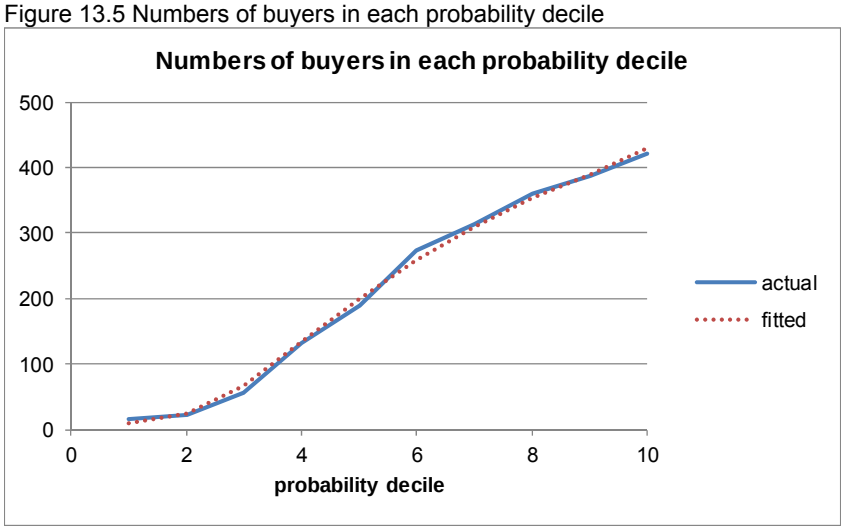
The logistic model in table 13.9 was derived from the 4,794 2000-01 SIHC households owned or being bought for which sale value was known. The probability of being bought increases strongly with income (presumably because many retired persons own their dwellings but have little income). Probability of being bought decreases strongly with the age of the household reference person. To balance approximately with the tenure proportions for each age group estimated for 09-10 by ABS (2011a table 14A), the adjustments in table 13.11 were made to the constant term of the model.

Table 13.9 Logistic model of probability of buying if buying or owning

Variable	Estimate	SE	P>|z|
Age	-0.06781	0.02259	0.003
Agesquare	-0.00042	0.00022	0.058
Income	0.210	0.040	0
Constant	4.152	0.559	0
N			4794
LLR			0.320

The 4,794 dwellings were sorted in order of their fitted probability of being bought, and grouped into deciles. Figure 13.5 shows the actual number of dwellings being bought in each decile, together with the sum of the fitted probabilities for the decile.

Figure 13.5 Numbers of buyers in each probability decile



13.12 Adjustments to model probabilities for moving households

The probability models in tables 13.7 to 13.9 represent the observed distributions of dwellings in 00-01, and are not directly applicable to moving households. Some behavior changes are likely to have occurred since 00-01. An iterative process was used to determine the model adjustments in table 13.10, with the adjustment term for each age for each new trial being adjusted by

ln (actual proportion / simulated proportion) / (1 – actual proportion)

Table 13.10 Adjustments to constants of probability models for moving households

Tenure	15-	25-	35-	45-	55-	65-	75-
Rent-free	-0.42	-1.37	-0.84	-0.16	0.05	0.28	0.23
Renting	0.86	-1.08	-0.48	-0.91	-0.75	-0.37	-0.91
Buying	37.19	1.91	0.84	1.01	0.28	-0.46	0.76

To smooth out random fluctuations, particularly for the small rent-free category, averages of five simulation runs were used. This form of iterative adjustment should give rapid convergence if the process being modelled is the result of a single logistic probability step. Housing tenures are however simulated as the results of household moves, which are likely to be frequent for young households. Transitions from buying to owned occur as mortgages are paid off. In spite of these complexities, the iterative process is reasonably robust, and only six iterations were needed to get the approximate matching shown in figures 13.6 to 13.9.

13.13 Simulated and actual tenures at 30/6/10

Figures 13.6 to 13.9 shows the simulated proportions with each type of house tenure, compared with survey data from SHIC 09-10. Although a good fit was obtained for each age range from 25-34 on, it was not possible to match the 0.6% ownership proportion for ages 15-24. Table 13.10 shows a large adjustment to the buying model for 15-24, so that no moving households are assumed to own without a mortgage. The simulated 4.9% ownership proportion at 15-24 is thus the result of dwellings owned outright at 30/6/01, or where a mortgage has been paid off after 30/6/01. Data adjustments are needed to get a better fit for the 15-24 group.

Figure 13.6 Percentages of dwellings owned without a mortgage

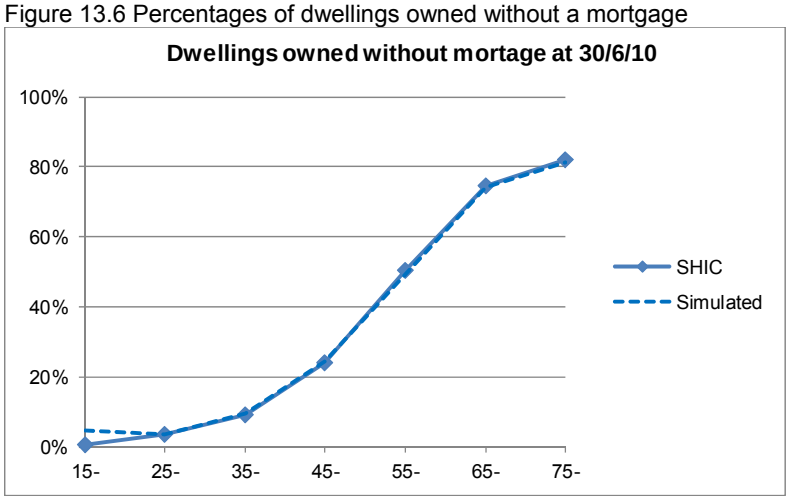
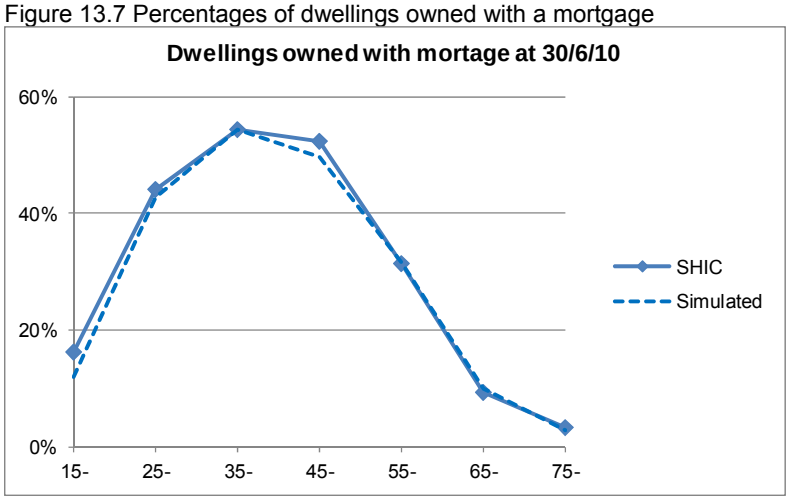
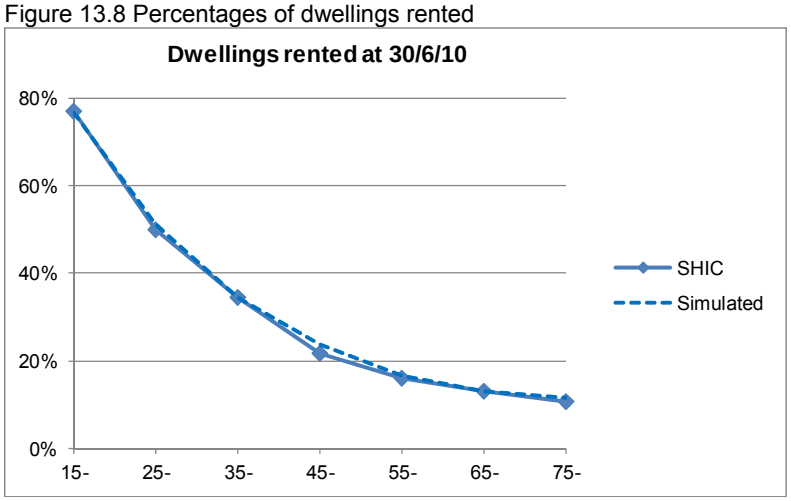


Figure 13.7 Percentages of dwellings owned with a mortgage



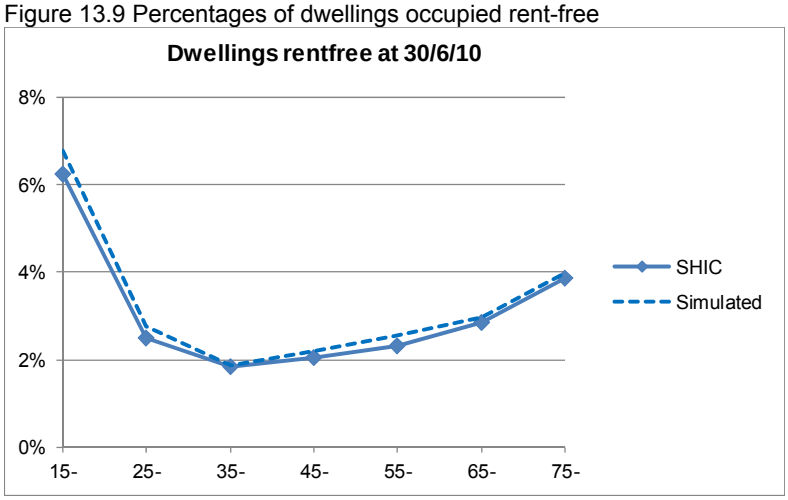
Although the fits are broadly reasonable, better fits could have been obtained with a few more iterations.

Figure 13.8 Percentages of dwellings rented



Reasonable fits were obtained for each age range.

Figure 13.9 Percentages of dwellings occupied rent-free

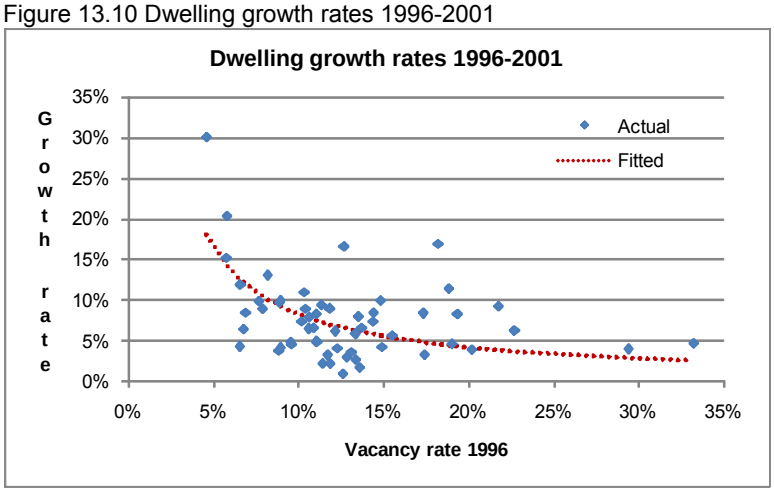


Young persons may occupy without paying rent dwellings owned by their parents. Some older persons may have life tenancies, for example where a widow has the right to occupy the family home, with the ownership reverting to the children on her death.

13.14 Private dwelling vacancies and growth rates

Figure 13.10 shows actual and fitted dwelling growth rates from 1996 to 2001, plotted against vacancy rates in 1996.

Figure 13.10 Dwelling growth rates 1996-2001



Dwelling vacancy rates at the 1996 and 2001 censuses, and growth rates in the total numbers of private dwellings, are from Australian Bureau of Statistics (2002, pp33-42). Actual and fitted growth rates are shown in figure 23.10 for 57 statistical divisions (counting the ACT as one division). Linear regression was used to fit the growth rate from 1996 to 2001 to the inverse of the vacancy rate at 1996, giving a fitted relationship of

$$\text{Growth rate 1996-2001} = 0.00825 / (\text{vacancy rate 1996}) \qquad (13.1)$$

The regression has an adjusted R-squared of 0.785, which is surprisingly high for what appears to be a poor fit.

13.15 Creating initial vacant dwellings, and subsequent new dwellings

The 2001 census sample file provided details of occupied private dwellings, but no details of vacant dwellings. The 2001 vacancy rates for each statistical division were assumed to apply, and initial vacant dwellings created by replicating randomly sampled occupied dwellings in the division. At the start of each subsequent projection year, the probability of any occupied dwelling being replicated as a vacant dwelling was assumed to be

.005 / region vacancy rate maximum of 0.2 (13.2)

At start of each projection year after the first, the probability of a vacant dwelling being demolished was assumed to be

(0.05 + 0.5 * region vacancy rate) * region vacancy rate (13.3)

The parameters if relationships 13.2 and 13.3 were determined by trial and error, trying to ensure that vacancy levels remained adequate for new households, but not excessive. The form and parameters for 13.3 are particularly uncertain.

13.16 Conclusions

The processes used to impute dwelling values to 2001 census sample data, and to correct apparent ownership mis-statements at young ages, are closely based on survey data (ABS 2003b). The modelling processes for moving households, and for periodic updating of mortgages and rents, appear reasonably robust. Assumptions made about the creation of new dwellings and the demolition of existing buildings are loosely based, but needed to ensure reasonable vacancy rates.

Assumptions made about tenure when households move into new dwelling, and about mortgage repayment rates, could be investigated using HILDA longitudinal data.

Although not yet attempted, the model's separate treatment of dwellings and households could allow simulations of buyer/seller interactions, and endogenous generation of market dwelling prices and rents. This could include separate modelling of dwellings owned as investments rather than for personal use.

14. DISEASE AND DISABILITY

National household microsimulation models have provided patchy coverage of disease and disability. Some have modelled disability statuses, without considering the underlying diseases responsible for the disabilities. Others have modelled the detailed progression of a few diseases.

The Cumpston model simulates the incidence and development of 123 diseases. These are based on the 169 diseases described in "The burden of disease and injury in Australia 2003" (Begg et al 2007), omitting those diseases with short durations or low disability levels.

Where possible, diseases are modelled by type and stage. For example, separate models are specified for three different types of breast cancer, showing the progression through four stages from primary therapy to terminal. In all, 584 disease stages are modelled, each with its own disability weight. Each projection cycle, the emergence of new diseases for each person is simulated, together with recovery from existing diseases, transition to a further stage, or death.

The disease stages for each person in the baseline file at 30/6/01 are simulated by assuming that current incidence, mortality, recovery and transition rates have remained unchanged in the past. Allowing for the "healthy immigrant" effect, immigrants are assumed to have had no diseases on arrival in Australia, but normal disease risks thereafter.

The housing and disease modelling in the Cumpston model could provide useful projections of the demand for aged care in local areas, and the potential ability of the aged to pay for their care.

14.1 Background

Late in 2009, the Australian government had asked its Productivity Commission to inquire into disability care and support, and into care for older persons. A decision was made in January 2010 to add detailed disease simulations to the Cumpston model, to assist submissions to the inquiries, and to assess the feasibility of disease modelling. This work was completed by July 2010, and three submissions made (Cumpston 2010 b, c). This chapter describes the modelling techniques and data sources, and draws some conclusions that may help other modellers.

14.2 Structure of Cumpston model

To provide the desired multi-disease simulations of disability and care, the Cumpston model was extended by

- Adding incidence, development, disability weight and mortality assumptions for 584 stages of 123 diseases likely to result in disability or death, based on models developed by the Australian Institute of Health and Welfare for their burden of disease and injury studies (Begg et al 2007)
- Using these assumptions to retrospectively impute up to five disease conditions for each person in the baseline file, with the disease with the lowest disability weight being dropped if necessary

- Projecting forward the incidence of new diseases, and the development of existing diseases, again allowing for up to 5 disease conditions for each person
- Simulating death separately for each disease condition for each person, allowing for the assumed mortality rate for that condition
- Using an approximate logistic model of aged care residence derived from SDAC 2003 unit records to retrospectively impute the care status of each person in the baseline file, and to simulate their admission to care during the projection.

14.3 Disease incidences, durations and prevalences

Table 14.1 shows summary details of 169 diseases, grouped into chapters. Incidence numbers for tables H to Q, S and Z are from annex table 15 to Begg et al (2007), and are estimates of the numbers of new incidents in 2003. Prevalence rates are from annex table 16, and are estimates of the numbers of persons with that condition at the middle of 2003. Durations were derived by dividing prevalences by incidences.

Table 14.1 Disease groups with AIHW incidence estimates

Chapter	Description	Diseases	Incidence m	Duration years	Prevalence m
A	Infectious and parasitic diseases	23	17.566	*0.12*	*2.091*
B	Acute respiratory infections	3	29.460	*0.01*	*0.419*
C	Maternal haemorrhage	3	0.072	*6.92*	*0.498*
D	Birth trauma and asphyxia	3	0.012	*11.09*	*0.131*
E	Nutritional deficiencies	3	0.943	*1.02*	*0.962*
F	Malignant neoplasms	26	0.470	*0.74*	*0.349*
G	Other neoplasms	2	0.021	*0.23*	*0.005*
H	Diabetes mellitus	2	0.097	12.07	1.171
I	Endocrine and metabolic disorders	3	0.001	52.23	0.028
J	Mental disorders	13	0.495	7.72	3.818
K	Nervous system and sense organ disorders	24	1.001	6.87	6.879
L	Cardiovascular disease	9	0.121	6.37	0.770
M	Chronic respiratory disease	2	0.099	17.63	1.744
N	Diseases of the digestive system	9	0.368	0.69	0.255
O	Genitourinary diseases	4	0.106	4.14	0.437
P	Skin diseases	4	0.200	2.59	0.519
Q	Musculoskeletal diseases	6	9.355	0.19	1.771
R	Congenital anomalies	10	0.004	*42.34*	*0.177*
S	Oral conditions	4	7.387	0.40	2.970
T	Unintentional injuries	12	0.290	*1.15*	*0.334*
U	Intentional injuries	3	0.041	*0.34*	*0.014*
Z	Ill-defined conditions	1	0.005	6.31	0.029
Total		169	68.113	0.37	25.369

Unfortunately, chapters A to G, R, T and U were not included in annex table 15. For these diseases, average duration estimates were obtained from the spreadsheets available on www.aihw.gov.au/bod as part of "The burden of disease and injury in Australia 1996". Prevalence estimates were then obtained by multiplying incidences by durations.

The average duration of infectious and parasitic diseases is about 6 weeks, with most of the order of a week, and a few more serious infections, such as HIV/AIDS and chronic hepatitis being much longer. Unintentional injuries have an average duration of 1.15 years, as they are based on accidents resulting in hospital admissions or

treatment in emergency departments. Data on intentional injuries are similarly based on hospital records.

14.4 Diseases and disease stages simulated

Many of the 169 diseases in table 14.1 can be subdivided by type or development stage. For example, the spreadsheets available as part of "The burden of disease and injury in Australia 1996" include 3 types of breast cancer, each with 4 development stages, as shown in table 14.2.

Table 14.2 Breast cancer types and stages

Type	Stage Number	Stage	Disability weight	Recovery constant	Transition constant
Tumour <2 cm	1	Diagnosis & primary treatment	0.26		4.62
Tumour <2 cm	2	Remission	0.26	0.29	0.10
Tumour <2 cm	3	Disseminated cancer	0.79		0.57
Tumour <2 cm	4	Terminal stage	0.93		
Tumour 2-5 cm	1	Diagnosis & primary treatment	0.69		2.83
Tumour 2-5 cm	2	Remission	0.26	0.28	0.23
Tumour 2-5 cm	3	Disseminated cancer	0.79		0.57
Tumour 2-5 cm	4	Terminal stage	0.93		
Tumour >5 cm	1	Diagnosis & primary treatment	0.81		1.50
Tumour >5 cm	2	Remission	0.26	0.23	0.55
Tumour >5 cm	3	Disseminated cancer	0.79		0.57
Tumour >5 cm	4	Terminal stage	0.93		

Disability weights are those assumed by Begg et al (2007 p17). They are intended to "quantify societal preferences for health states in relation to the societal idea of good health". Other examples are 0.07 for diabetes, 0.125 for a slipped disc with chronic pain, 0.27 for mild dementia, 0.43 for blindness, 0.57 for paraplegia and 0.76 for unremitting unipolar major depression. These disability weights are inherently subjective, and may not be appropriate as admission criteria for disability benefits.

The only pathways assumed from each disease stage are recovery, transition to the next stage of the disease, and death. Probabilities of each of these events occurring for each event for each disease stage are assumed to be constants depending only on sex and age. If f(x) is the probability of being in a particular disease stage at time x, given the person is in that state at time 0, then

$$df(x)/dx = - (\lambda_m + \lambda_r + \lambda_t) f(x) \qquad (15.1)$$

where λ_m is the mortality constant, λ_r the recovery constant and λ_t the transition constant. With this notation

$$f(x) = \exp[- (\lambda_m + \lambda_r + \lambda_t) x] \qquad (15.2)$$

Mortality, recovery and transition constants were estimated a range of information in the "The burden of disease and injury in Australia 1996" spreadsheets, including average stage durations, relative mortality rates and 5-year survival rates.

In all, 123 of the 169 diseases were chosen for simulation, with the omitted diseases being those with durations under 6 months or very low disability weights. These 123 selected diseases were identified as having a total of 584 stages. For all but one of these stages, disability weights and recovery, transition and mortality parameters were estimated from the spreadsheets. The exception was a one-stage disease called "back problems", where the spreadsheet assumed high incidences, each with an average duration of 0.011 years. Incidence rates for chronic back pain, assuming no

recovery or extra mortality, were estimated from the unit record data for the Survey of Disability and Carers 2003 (ABS 2005).

14.5 Continuous time disease simulation within projection periods

Some of the average durations for disease stages are very short. For example, the average duration assumed for the terminal stages of most cancers is one month. To allow for short durations without excessive calculation times, continuous time simulation is used for disease events within projection periods. From equation 15.2, the time x since entering a disease state is

$$x = - \ln(f(x)) \ / \ (\lambda_m + \lambda_r + \lambda_t) \qquad (15.3)$$

For example, a person in remission from breast cancer with tumours under 2 cm has a recovery time constant of 0.29, and a transition time constant of 0.10. The mortality constant is zero, as no deaths from cancer are assumed until the terminal stage. The time until either remission or transition occurs is simulated by selecting a random number r between 0 and 1, and calculating the time as - ln(r) / 0.39. If this simulated time is less than the time remaining to the end of the simulation period, an event is assumed to occur. If more than one type of event is possible, then the choice between event types is made by selecting another random number, and choosing in proportion to the time constants of the possible events.

14.6 Constructing base diseases in 2001 using AIHW disease models

The initial diseases and disease stages for each of the 175,044 persons in the baseline data at 30/6/01 were simulated by an iterative process

- For persons born in Australia, the occurrence of a congenital or birth-related defect was simulated
- Stepping forward a year at a time from their birth or immigration date, the occurrence of new diseases, and the development of existing diseases, was simulated
- Each year, death from each disease was simulated, taking into account only the extra risks of death from disease
- If death from disease occurred, the process was restarted at the birth or immigration date.

This process assumes that the age-specific incidence, recovery, transition and mortality risks from each disease have remained unchanged up to 2001. The Australian Bureau of Statistics (2004b p3) noted that there was “little change in the disability rate between 1998 (20.1%) and 2003 (20.0%)”. Begg et al (2007 p33-34) noted the considerable increase in the incidence of type 2 diabetes, and the lack of any data suggesting trends in mental health, hearing loss, vision loss and musculoskeletal disorders.

The above process also assumes that immigrants come to Australia with no diseases. Kennedy, McDonald and Biddle (2007) provide evidence of the “healthy immigrant effect” in the US, Canada, UK and Australia. For Australia, they show that the incidence of chronic conditions is substantially lower for all immigrant regions, self-assessed health generally better, and obesity and smoking rates lower. Chronic hepatitis B prevalence rates are however higher in some immigrants, particularly those from the Asia-Pacific region (Butler, Korda, Watson & Watson 2009 p11). More exact allowances for immigration effects could be included in the above process.

The above process assumes that existing diseases are uncorrelated with a person's employment and household status. In practice, however, persons with severe disabilities are less likely to be employed, and less likely to be in partnerships or living in private dwellings. A different method of constructing baseline diseases is needed. Further work is planned on disease imputation methods.

14.7 Comparing simulated diseases with 2003 survey data

Table 14.3 compares the numbers of each disability estimated from the unit records from the 2003 Survey of Disability and Carers (ABS 2005) with those imputed at 2001, and then projected forward two years allowing for births, deaths, immigration and emigration, and for disease incidence and development. As in the baseline projection to 2001, immigrants are assumed to come to Australia with no diseases.

Table 14.3 Observed and simulated numbers of persons with main conditions in 2003

ICD10 chapter	Description of main condition	Simulated m	SDAC m
I	Infectious & parasitic diseases	0.015	0.030
II	Neoplasms	0.182	0.103
III	Diseases of the blood & blood-forming organs	0.426	0.018
IV	Endocrine, nutritional & metabolic diseases	0.294	0.471
V	Mental & behaviourial disorders	3.547	0.912
VI	Diseases of the nervous system	0.145	0.521
VII	Diseases of the eye & adnexa	0.216	0.108
VIII	Diseases of the ear & mastoid process	0.725	0.454
IX	Diseases of the circulatory system	0.195	1.010
X	Diseases of the respiratory system	0.809	1.068
XI	Diseases of the digestive system	0.054	0.159
XII	Diseases of the skin & subcutaneous tissue	0.321	0.067
XIII	Musculoskeletal	1.226	2.348
XIV	Diseases of the genitourinary system	0.198	0.072
XVI	Certain conditions originating in the perinatal period	0.010	0.003
XVII	Congenital malformations	0.017	0.049
XVIII	Symptoms, signs and abnormal findings, nec	0.000	0.182
XX	External causes of morbidity & mortality	0.999	0.512
Total		9.380	8.087

The main condition for each simulated person was taken as the disease with the highest AIHW disability weight. The simulated numbers were multiplied by 19.719/ 0.178 (the ratio of the Australian population at 30/6/03 to the simulated numbers of persons at 30/6/03).

There are many reasons to expect differences between the survey and simulated numbers:

- Random variations, both in the simulations and the survey data
- Some persons may not have reported certain conditions because of the sensitive nature of the condition - eg alcohol and drug-related conditions, schizophrenia, mental retardation or mental degeneration (ABS 2004 60)
- Some conditions may be episodic or seasonal - eg asthma, epilepsy
- Lack of awareness of the presence of a condition, such as mild diabetes
- Lack of comprehensive medical information kept by cared-accommodation establishments, who completed survey returns on behalf of their residents

- The simulations rest on incidence, recovery, transition and mortality assumptions separately derived for each of 123 diseases, together with the assumption that these rates have not changed in the past.

14.8 Simulated and actual deaths in 2002

The main causes of death in 2002 were estimated by projecting for two years from 30/6/01. These gave 2618 deaths, and this low number means that there are substantial uncertainties in the simulated numbers of deaths, particularly for the less common causes. Table 14.4 compares their composition with the main causes of death reported for Australia in 2002 (Australian Bureau of Statistics 2010c).

Table 14.4 Simulated causes of death in 2002, compared with actual

ICD10 chapter	Description of main cause of death	Simulated deaths	Actual deaths
I	Infectious & parasitic diseases	1.1%	1.3%
II	Neoplasms	25.6%	28.7%
III	Diseases of the blood & blood-forming organs	0.1%	0.3%
IV	Endocrine, nutritional & metabolic diseases	6.7%	3.5%
V	Mental & behaviourial disorders	1.3%	2.4%
VI	Diseases of the nervous system	18.0%	3.5%
VII	Diseases of the eye & adnexa	0.0%	0.0%
VIII	Diseases of the ear & mastoid process	0.0%	0.0%
IX	Diseases of the circulatory system	31.4%	37.6%
X	Diseases of the respiratory system	7.9%	8.7%
XI	Diseases of the digestive system	0.7%	3.3%
XII	Diseases of the skin & subcutaneous tissue	0.0%	0.2%
XIII	Musculoskeletal	0.3%	0.8%
XIV	Diseases of the genitourinary system	1.0%	2.2%
XV	Pregnancy, childbirth & the puerperium	0.0%	0.0%
XVI	Certain conditions originating in the perinatal period	0.1%	0.5%
XVII	Congenital malformations	0.0%	0.4%
XVIII	Symptoms, signs and abnormal findings, nec	0.0%	0.6%
XX	External causes of morbidity & mortality	5.9%	5.8%
Total		100.0%	100.0%

Some of the possible causes of the differences between simulated and actual deaths are

- Random variations in the simulations and the actual experience
- Increases in assumed mortality rates for some deaths of the nervous system, made in an effort to match the high mortality rates of persons in aged care in Australia
- Uncertainty about the mortality assumptions for some diseases in the AIHW disease models
- Reluctance by medical practitioners to show dementia as the main cause of death in death certificates.

The AIHW disease models omit immediate deaths from external causes, so a “fatal external cause” disease subtype was added, with mortality assumptions for each sex and age-group chosen so as to give simulations matching the observed 2002 deaths from external causes.

14.9 Models of the probability of being in aged care

The logistic models in table 14.5 were fitted to 39,086 persons without dementia, and 2,147 persons with dementia, in the unit records for SDAC 2003 (ABS 2005). The models were fitted by weighted stepwise backwards regression, omitting variables with a significance probability of 10% or more. The log likelihood reduction was 0.538 for the without dementia model, but only 0.105 for the with dementia model.

Table 14.5 Logistic models of the probability of being in aged care, without & with dementia

Variable	Without dementia	Without dementia	With dementia	With dementia
	Coefficient	SE	Coefficient	SE
age	0.143	0.004	0.058	0.014
cerebral palsy	2.854	0.669		
congenital	1.527	0.369		
depression	2.145	0.142		
diabetes	0.416	0.116		
epilepsy	1.632	0.337	1.422	0.449
head injury			0.814	0.280
heart	0.755	0.105	0.603	0.246
married	-1.693	0.100	-0.647	0.220
multiple sclerosis	4.963	0.476		
neoplasm	0.686	0.175	0.623	0.373
paralysis	3.727	0.380	3.130	1.098
parkinson	2.737	0.299	1.480	0.411
retardation	2.099	0.385		
schizophrenia	3.508	0.355	1.642	0.532
sex			0.417	0.215
stroke	0.875	0.118		
urinary	1.204	0.312		
constant	-15.119	0.346	-5.129	1.202
N		39086		2147
log likelihood reduction		0.528		0.105
adjustment to constant	-0.800		-0.200	

Dementia is a key determinant of the probability of being in aged care - 0.4% of persons without dementia are in aged care, and 64% of those with dementia are in aged care. The probability of being in aged care increases linearly with age. For those without dementia, married persons have a much lower probability of being in aged care.

14.10 Simulated and actual persons in aged care in 2003

In the absence of longitudinal data on aged care residence, probabilities of admission to aged care were estimated from the models in table 14.5, adjusting the regression constants by the amounts in the last line of the table. These adjustments were chosen so as to simulate approximately correct numbers of persons in aged care in 2003, separately for persons without and with dementia. Given the data limitations, the results in table 14.6 seem broadly reasonable.

Table 14.6 Characteristics of simulated and actual persons in care in 2003

Personal characteristics	Simulated	Actual	Source
Average age of persons in care	85.6	83.4	AIHW 2004 table 2.2
% of all persons in care	0.71%	0.72%	AIHW 2004 table 2.2
% in care who are male	33.4%	27.8%	AIHW 2004 table 2.2
% in care with dementia	42.6%	44.5%	SDAC 2003
Admissions to care pa, as % of those in care	34.6%	37.0%	AIHW 2004 table 3.2
Deaths in care pa, as % of those in care	29.5%	28.7%	AIHW 2004 table 3.5

A major problem is that the assumed subdivisions of disease types by disability weight, largely based on AIHW disease models, do not give adequate differentiation between mortality rates. The assessment procedures used for admission to aged care have clearly been able to select persons with very high mortality rates. More data on the progression of mortality rates are needed, particularly for dementia.

From 20/3/08, the Aged Care Funding Instrument (ACFI) replaced the Residential Classification Scale used for determining care payments for residents in aged care homes (AIHW 2009 p136). The mental, behavioural and medical diagnoses made as part of the ACFI assessments could provide valuable data to directly model admission probabilities.

14.11 Existing disease and disability models

Spielauer (2007) mentions seven microsimulation models which have simulated diseases or disabilities

- DYNASIM3 uses a discrete-time hazard model to simulate entering, exiting or remaining in a work-limitations disability status (Favreault & Smith (2004 p10)
- CORSIM modelled four risk factors – smoking, alcohol consumption, sugar consumption and diabetes, as well as disability status
- DYNAMOD modelled disability, affecting mortality and education
- LifePaths models five disability statuses, developed specifically for the analysis of future care needs
- POHEM (Population Health Model) models weight gain, smoking, total cholesterol, blood pressure and alcohol consumption as risk factors, with diseases modelled including breast, lung and colorectal cancers and acute myocardial infarction
- MOSART models a limited set of health-related behaviours, including moving in and out of old age care institutions, disability and rehabilitation
- NCSSU models long-term care strategies for persons aged 65, by linking with the PRSSU cell-based long-term care macro model.

APPSIM was originally intended to model disability levels, but not risk factors or diseases (Kelly 2007 p16). The aims for the health module were refined to providing projections of health care expenditure, with the initial sub-modules being health service

usage (medical, hospital and pharmaceutical) and health care expenditure (Lymer 2009 p29). Self-assessed health status was to be trialled as a health status indicator.

Walker, Butler & Calagiuri (2010) describe the development of an "umbrella" model representing the full Australian population, together with submodels for selected chronic diseases. The umbrella model will contain details of any existing chronic diseases, together with risk factors such as body mass index, self-rated health, socio-economic status, work status, smoking, alcohol use, blood pressure, cholesterol and blood sugar levels. The initial submodels are for diabetes and its main complications, and cardiovascular disease. The overall model is intended to help assess the benefits and costs of potential chronic disease interventions. It uses as base data the 25,906 weighted unit record data from the Australian National Health Survey 2004-2005, and reweights each 5 years so as to align with ABS population projections.

14.12 Getting better data

The Cumpston model uses many assumptions derived from the HILDA longitudinal survey (Watson 2008), but HILDA does not record details of individual diseases. Walker, Butler & Colagiuri (2010) used data from the 2000 and 2005 waves of the Australian Diabetes, Obesity and Lifestyle Survey. The Productivity Commission draft report on disability care and support used data from the Multiple Sclerosis Longitudinal Study (2011 p14.4). AIHW used statistical and anecdotal data from many different sources to construct disease models (Begg et al 2007). Longitudinal studies however suffer from a range of problems, including inaccurate reporting, withdrawals, privacy restrictions and inadequate panel sizes.

As much use as possible thus needs to be made of records maintained for administrative purposes. At least in Australia, statistical access to such records has recently become easier, subject to reasonable privacy restrictions. Some possible data sources for a comprehensive disease model include

- ACFI assessments for persons entering residential care, or being reassessed while in care
- Deaths of persons in residential care
- Claims under the Pharmaceutical Benefits Scheme
- Claims under the Medicare Benefits Schedule.

14.13 Conclusions

The Cumpston model was extended to include simulation of 123 disease types. Conclusions from this were

- Valuable data are potentially available from present administrative data records
- Present AIHW disease models do not give adequate mortality rate differentiation
- Expert medical advice is needed to help derive assumptions.
- Given data, risk factors and interactions between diseases could be added
- The Cumpston model is robust enough to support major changes to its functionality.

15. CONCLUSIONS AND FURTHER WORK

This thesis set out to find new methods for household microsimulation, and to demonstrate their application to Australian data. It proposes faster methods to sample, align and match, with equal or better fidelity than existing methods. The Cumpston model, based on a 1% sample of Australian households, was created to test these suggestions and explore the feasibility of larger models. A 10% model of Australian households, using techniques tested in the Cumpston model, is proposed.

15.1 Sampling methods

All-case simulation, where demographic outcomes are simulated for each person each year in file order, has been used in almost every national household microsimulation model. This thesis proposes sampling with loaded probabilities, a form of stratified sampling, where the probabilities for all persons in a strata are loaded so that no probability exceeds unity. Tests with the Cumpston model show sampling with loaded probabilities can reduce run times of demographic simulations with yearly cycles by about 43%, and run times with weekly cycles by about 98%.

15.2 Event alignment methods

This thesis proposes proportional event alignment, as a modest improvement to Johnson's two-stage alignment by sorting (2001). It also proposes event alignment using random selection, a one-stage method successfully implemented in the Cumpston model. Regardless of the choice of alignment method, care is needed In the choice of alignment pools.

15.3 Matching methods

The model described by Orcutt et al (1961) immediately found an appropriate partner for any person simulated as marrying. Most subsequent models have used a batch process, where males and females are separately selected for marriage, and matches formed subsequently. This thesis proposes immediate probability-weighted matching and immediate "best of n" matching, and uses data on Australian marriages to show that these methods give better results than several batch matching methods. The Cumpston model uses best of n matching, and closely replicates the means and standard deviations of age differences for 5 types of household relationship (see 9.19),

15.4 Data sources

Baseline data for the Cumpston model came from the 2001 1% census sample file, supplemented by the Survey of Income & Housing Costs 00-01 and the HILDA longitudinal survey. Administrative data were used to approximately correct some serious biases in the census data, particularly education levels,. Administrative data were also very useful in supplementing the HILDA data.

15.5 Validation of Cumpston model

A wide range of validation techniques were used.to ensure that the model was reasonably close to reality in each of its processes (see 7.17). Output of time series into a single Excel workbook, providing a comprehensive set of comparative graphs, proved particularly useful. Adjustments to some projection assumptions were needed to match the large changes in the decade since 2001.

15.6 Proposed 10% model of Australian households

The Cumpston model has shown that a 10% model of Australian households, with 2200 geographic subdivisions, is feasible. Such a model might take about 10 minutes for a 50 year simulation, assuming use of multithreading where feasible. Baseline data for such a model would have to be synthesized from census marginal totals, using combinatorial optimization or sequential allocation.

BIBLIOGRAPHY

Abello A, Kelly S & King A (2002) "Demographic projections with DYNAMOD-2", National Centre for Social & Economic Modelling, Technical Paper no. 21, May, vi + 50 pages

Ahmed V & O'Donoghue C (2007) "CGE-microsimulation modelling: a survey", Munich Personal RePEc Archive paper no 9307, posted 25/6/08, http:/mpra.ub.uni-muenchen.de/9307, downloaded 11/11/09

Albahari J & Albahari B (2008) "C# 3.0 pocket reference", O'Reilly, Sebastapol CA, xi + 230 pages

Albert C, Berteau-Rapin C & Di Porto A (2009) "PRISME: A dynamic microsimulation model for the French pension scheme CNAV", paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 33 pages

Albert G (2009) personal communication July 9

Anderson JM (1997) "CORSIM", Society of Actuaries, Schaumburg, Illinois, chapter 5 of a review of microsimulation models, www.soa.org

ABS (2002) "Census of population and housing - selected social and housing characteristics Australia 2001", catalog no 2015.0, Canberra June 17, iii + 130 pages

ABS (2003a) "Census of population and housing - selected education and labour force characteristics Australia 2001", catalog no 2017.0, Canberra February 18, v + 148 pages

ABS (2003b) "Survey of income and housing costs Australia 2000-01 - confidentialised unit record file (CURF) technical paper", catalog no, 3222.0, Canberra August, 63 pages

ABS (2003c) "Population projections Australia 2002-2101", catalog no 3222.0, Canberra September 3, 186 pages

ABS (2003d) "Census of population and housing - housing sample file Australia 2001", technical paper, catalog no, 3027.0, Canberra September 29, v + 57 pages

ABS (2003e) "Marriages and divorces Australia 2002", catalog no 3310.0, Canberra November 6, 92 pages

ABS (2003f) "Deaths Australia 2002", catalog no 3302.0, Canberra December 2, 111 pages

ABS (2004a) "Household and family projections, Australia 2001 to 2026", Canberra 18/6/04, iii + 138 pages

ABS (2004b) "Disability, ageing and carers: summary of findings Australia 2003", catalog no 4430.0, Canberra 15/9/04, 79 pages

ABS (2005) " Basic confidentialised unit record file: survey of disability, ageing and carers 2003 (reissue)", Canberra July 22, iii + 200 pages

ABS (2008) "Population projections Australia 2006 to 2101", catalog no 3222.0, Canberra September 4, downloaded from www.abs.gov.au/statistics

ABS (2010a) "Schools Australia 2009 - NSSC table 40a & 41b", catalog no, 6523.0, Canberra, downloaded from www.abs.gov.au/statistics 21/3/11

ABS (2010b) "Australian demographic statistics" catalog no 3101.0, Canberra, downloaded from www.abs.gov.au/statistics 21/3/11

ABS (2010c) "Deaths Australia 2009" catalog no 3302.0, Canberra, downloaded from www.abs.gov.au/statistics 8/11/11

ABS (2011a "Household income and income distribution, Australia - Detailed tables 2009-10" catalog no 6523.0, Canberra, August 31, downloaded from www.abs.gov.au/statistics 23/9/11

ABS (2011b) "House price indexes: Eight capital cities" catalog no 641.0, Canberra, August 31, downloaded from www.abs.gov.au/statistics 20/9/11

ABS (2011c) "Births Australia 2010" catalog no 3301.0 Canberra, October 25, downloaded from www.abs.gov.au/statistics 8/11/11

ABS (2011d) "Overseas arrivals and departures, Australia 2010" catalog no 3401.0 Canberra, downloaded from www.abs.gov.au/statistics 9/11/11

ABS (2011e) "Labour force, Australia detailed quarterly" catalog no 6291.0.55.003 Canberra, downloaded from www.abs.gov.au/statistics 9/12/11

ABS (2011f) "Schools Australia 2010 - NSSC table 40a & 41b", catalog no, 6523.0, Canberra, downloaded from www.abs.gov.au/statistics 6/12/11

AIHW (2004) "Residential aged care in Australia 2002-03 - a statistical overview", xi + 90, downloaded from www.aihw.gov.au

AIHW (2009) "Residential aged care in Australia 2007-08 - a statistical overview", June, viii + 151, downloaded from www.aihw.gov.au

APRA (2005) "Superannuation trends September 2004", 11/1/05, downloaded from www.apra.gov.au 25/7/11

APRA (2007) "Insight - celebrating 10 years of superannuation data collection 1996-2006", Sydney, issue, 65 pages, downloaded from www.apra.gov.au 21/10/09

APRA (2011a) "Annual superannuation bulletin, June 2009", January 19, downloaded from www.apra.gov.au 11/7/11

APRA (2011b) "Annual superannuation bulletin, June 2010", January 19, downloaded from www.apra.gov.au 11/4/11

ATO (2011a), benefits, contributions, rollovers, investment earnings and final balances for each member of a self-managed superannuation fund in 07-08 and 08-09, subdivided by sex and 5-year age group, supplied in response to a Freedom of Information request, June 29

ATO (2011b) "Key superannuation rates and thresholds", downloaded from www.ato.gov.au 27/7/11

Baekgaard, H (2002) "Micro-macro linkage and the alignment of transition processes - some issues, techniques and ", National Centre for Social and Economic Modelling, Canberra, Technical Paper no 25, June, v + 29 pages

Bacon B & Pennec S (2007) “APPSIM - modelling family formation and dissolution”, National Centre for Social and Economic Modelling, Canberra, Working Paper no 4, November, v + 29 pages

Begg S, Vos T, Barker B, Stevenson C, Stanley L & Lopez A (2007)“The burden of disease and injury in Australia 2003”, Australian Institute of Health and Welfare, May, 136 pages plus appendices

Birkin M & Clarke G (2009) “Improving the representation of diversity in spatial microsimulation models", presented to the 2009 International Microsimulation Association Conference, Ottawa June 8-10, available from http://www.microsimulation.org/IMA/Ottawa_2009.htm

Birkin (2009) personal communication, November 18

Blanchet D, Buffeteau S, Crenner E & Le Minez S (2009) “The Destinie 2 microsimulation model: overview and illustrative results”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 22 pages, available from http://www.microsimulation.org/IMA/Ottawa_2009.htm

Bonnet C & Mahieu R (1998) “Public pensions in a dynamic microanalytic framework: the case of France”, in “Microsimulation modelling for policy analysis - challenges and innovations”, eds Mitton S, Sutherland H & Weeks M, Cambridge University Press, 2000 p175-199

Bouffard N, Easther R, Johnson T, Morrison RJ & Vink J (2001) “Matchmaker, matchmaker, make me a match”, Joint CORSIM and DYNACAN Project Document, September, 19 pages (also in Brazilian Electronic Journal of Economics, vol 4, no 2)

Butler JRG, Korda RJ, Watson KJR & Watson DAR (2009) “The impact of chronic hepatitis B in Australia: Projecting mortality, morbidity and economic impact”, Australian Centre for Economic Research on Health research report no 7, September, viii +66 pages

Caldwell SB (1983) "Broadening policy models: alternative strategies", in "Microanalytic simulation models to support social and financial policy", eds Orcutt GH, Merz J & Quinke H, North-Holland, Amsterdam 1986, p59-77

Caldwell SB (1993) “The CORSIM dynamic microsimulation model”, in “Microsimulation and public policy”, ed Harding A, North Holland, Amsterdam,1996, p505-522

Caldwell SB & Morrison RJ (1998) “Validation of longitudinal dynamic microsimulation models: experience with CORSIM and DYNACAN”, in Mitton L, Sutherland H & Weeks L (eds) “Microsimulation in modelling for policy analysis: challenges and innovations”, Cambridge, Cambridge University Press, p200-225

Charette C & Gribble S (2009) “Modelling large interacting populations in continuous time: a novel computational approach using Modgen”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 10 pages, available from http://www.microsimulation.org/IMA/Ottawa_2009.htm

Chenard D (1998) “Individual alignment and group processing: an application to migration processes in DYNACAN”, in Mitton L, Sutherland H & Weeks L (eds) “Microsimulation in modelling for policy analysis: challenges and innovations”, Cambridge, Cambridge University Press, p238-247

Clarke P & Leigh A (2011) “Death, dollars and degrees: socio-economic status and longevity in Australia”, Economic Papers, v 30, no 3, September p348-355

Cumpston JR & Sarjeant HB (1998) “Models of migration between Australian states and statistical divisions”, paper presented to the 9th national conference of the Australian Population Association, 29/9/98-2/10/98, Brisbane, 23 pages

Cumpston JR (2001) “Past and future causes of death - implications for planners”, paper presented to the National Annual Scientific Conference of the Australian Association of Gerontologists, Canberra 5-7 September 2001, 17 pages

Cumpston JR (2002) "Population projections 1996 to 2051 - methods, assumptions and results", Cumpston Sarjeant Pty Ltd report to CSIRO, 22 January, 5 pages + result files

Cumpston JR (2007) “Dynamic microsimulations of Australian individuals and households at varying geographic scales”, paper presented to the first general conference of the International Microsimulation Association, Vienna, August, 7 pages

Cumpston JR (2009a), “Acceleration, alignment and matching in multi-purpose household microsimulations”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 22 pages, available from http://www.microsimulation.org/IMA/Ottawa_2009.htm

Cumpston JR (2009b) “Simulation by sampling with loaded probabilities”, paper submitted for publication to the International Journal of Microsimulation

Cumpston JR (2010a) “Alignment and matching in multi-purpose household microsimulations”, International Journal of Microsimulation 3(2) p34-45

Cumpston JR (2010b), submissions no 7 and 78 to the Productivity Commission inquiry into disability care and support, available on www.pc.gov.au/projects/inquiry/disability-support

Cumpston JR (2010c), submission no 255 to the Productivity Commission inquiry into aged care, available on www.pc.gov.au/projects/inquiry/aged-care

Cumpston JR (2011) “An Australian disease and long-term care microsimulation model” paper presented to the third general conference of the International Microsimulation Association, Stockholm, June 8-10, 22 pages, available from http://www.scb.se/Grupp/Produkter_Tjanster/Kurser/_Dokument/IMA/Cumpston_paper_disease_May_10_2011.pdf

Dekkers G, Buslei H, Cozzlino M, Desmet R, Geyer J, Hofmann D, Raitano M, Steiner V, Tanda P, Tedeschi S, Verschueren F (2009) “What are the consequences of the AWG-projections for the adequacy of social security pensions?”, ENEPRI Research Report No 65 AIM WP4, January, 340 pages (available from www.enepri.org or www.ceps.eu)

Dekkers G (2009) personal communication December 31

Dekkers G (2010) personal communication December 8

Dekkers G (2011) personal communication October 15

Dekkers G & Cumpston JR (2011) “On weights in dynamic-ageing microsimulation models”, presented at the third general conference of the International Microsimulation Association, Stockholm June 8-10, available on http://www.scb.se/Grupp/Produkter_Tjanster/Kurser/_Dokument/IMA/Dekkers_cumpston.pdf

Department of Education, Employment and Workplace Relations (2010a) “Students 2009 full year selected higher education statistics”, downloaded from www.deewr.gov.au/HigherEducation/Publications/HEStatistics on 13/12/10

Department of Education, Employment and Workplace Relations (2010b) “Award course completions 2009”, downloaded from www.deewrv.au/HigherEducation/Publications/HEStatistics on 13/12/10

Department of Education, Employment and Workplace Relations (2011a) “Students 2009 full year selected higher education statistics”, downloaded from www.deewr.gov.au/HigherEducation/Publications/HEStatistics on 6/12/11

Department of Education, Employment and Workplace Relations (2011b) “Award course completions 2009”, downloaded from www.deewrv.au/HigherEducation/Publications/HEStatistics on 6/12/11

Easther R & Vink J (2000) “A stochastic marriage market for CORSIM”, Strategic Forecasting Technical Report, October, 8 pages

Emmerson C, Reed H & Shephard A (2004) “An assessment of Pensim2”, The Institute for Fiscal Studies, London, WP04/21, 51 pages

Evandrou M, Falkingham J, Johnson P, Scott A & Zaidi A (2003) “The SAGE model: a dynamic microsimulation population model for Britain”, in “Modelling our future - population ageing health and aged care”, eds Gupta A & Harding A, Elseveir, Amsterdam, 2007, xxv + 566 pages

Farooq B, Salvini PA & Miller EJ (2008) “Development of an operational integrated urban model system - volume X - ILUTE software documentation”, University of Toronto, December, 50 pages

Favreault M & Smith K (2004) “A primer on the dynamic simulation of income model (DYNASIM3)”, The Urban Institute, Washington DC, February 2004, 22 pages

Flood L, Jansson F, Pettersson T, Sundberg O & Westerberg A (2005) “SESIM III: A Swedish dynamic micro simulation model”, Sweden, Ministry of Finance, December 22, 54 pages

Fredricksen D, Knudsen P & Stolen NM (2009) “Updating and extending the data base for the dynamic microsimulation model MOSART”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 11 pages

Fredricksen D, Knudsen P & Stolen NM (2011) “The dynamic cross-sectional microsimulation model MOSART”, paper presented to the third general conference of the International Microsimulation Association, Stockholm, June 8-10, 8 pages

Galler HP (1992) “Competing risks and unobserved heterogeneity, with special reference to dynamic microsimulation models”, in “Household demography and household modelling”, eds van Imhoff E, Kuijsten A, Hooimeijer & van Wissen L (1995) Plenum Press New York & London, p203-224

Galler HP (1997) “Discrete time and continuous-time approaches to dynamic microsimulation reconsidered”, National Centre for Social & Economic Modelling, Canberra, Technical Paper No 13, October, v + 35 pages

Galler HP & Wagner G (1983) "The microsimulation model of the Sfb 3 for the analysis of economic and social policies", in "Microanalytic simulation models to support social and financial policy", eds Orcutt GH, Merz J & Quinke H, North-Holland, Amsterdam 1986, p227-247

Gardner J & Oswald S (2004) “How is mortality affected by money, marriage and stress?”, Journal of Health Economics

Gorgens T & Ryan C (2005) “A bounds analysis of school completion rates in Australia”, Research School of Social Sciences, Australian National University, April, 30 pages, downloaded from http://econrsss.anu.edu.au/Staff/ryan/pdf/AUbounds2.pdf on 6/10/09

Gunnes T (2009) “Education projections in Norway”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 15 pages

Gupta A & Harding A (2007) (eds) “Modelling our future: population ageing, health and aged care” International Symposia in Economic Theory and Econometrics, North-Holland, Amsterdam

Hagerstrand T (1957) “Migration and area: survey of a sample of Swedish migration fields and hypothetical considerations on their genesis”, Lund Studies in Geography, no 13 p27-158

Harding A (2007) “Challenges and opportunities of dynamic microsimulation modelling”, plenary paper presented to the first general conference of the International Microsimulation Association, Vienna, August 21, 23 pages

Harding, Keegan & Kelly (2010) “Validating a dynamic population simulation model: recent experience in Australia”, International Journal of Microsimulation 3(2) p46-64

Holm E, Holme K, Makila K, Mattsson-Kauppi M, Mortvik G (2002) “The SVERIGE spatial microsimulation model - content, validation, and example application”, Spatial Modelling Centre, Kiruna, Sweden, 74 pages (available on www.umu.se)

Hosmer DW & Lemeshow S (2000) “Applied logistic regression”, 2nd edition, Wiley Interscience, New York, xii + 375 pages

Inagaki S (2009a) “INAHSIM: A Japanese microsimulation model”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 15 pages

Inagaki S (2009b) “Effects of proposals for pension reform on the income distribution of the elderly in Japan”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 20 pages

Inagaki S (2009c) personal communication September 11

Inagaki S (2011) “Simulating policy alternatives to public pensions in Japan”, presented at the third general conference of the International Microsimulation Association, Stockholm June 8-10, available on http://www.scb.se/Grupp/Produkter_Tjanster/Kurser/_Dokument/IMA/inagaki_paper_rev1.pdf

International Microsimulation Association (2009) http//www.microsimulation.org

Jansen RPS & Dill SK (2009) “When and how to welcome government into the bedroom”, Medical Journal of Australia 190(5) p232-233

Johansen L (1960) “A multi-sectorial study of economic growth”, North-Holland, Amsterdam

Johnson T (2001) “Nonlinear alignment by sorting”, CORSIM Working Paper, February

Keegan M (2009) “Mandatory superannuation and self-sufficiency in retirement - an application of the APPSIM dynamic microsimulation model”, paper prepared for the Australian Conference of Economists 2009, Adelaide 28-30 September, 33 pages

Keegan M & Thurecht L (2008) “APPSIM - modelling of earnings”, National Centre for Social & Economic Modelling, Canberra, Working Paper No 10, June, v + 46 pages

Kelly S (2007) “APPSIM - objectives and requirements”, National Centre for Social & Economic Modelling, Canberra, Working Paper No 1, April, iv + 19 pages

Kelly S (2010) personal communication December 6

Kelly S & Keegan M (2008) “APPSIM – modelling household savings”, National Centre for Social & Economic Modelling, Canberra, Working Paper No 9, May, vi + 41

Kelly S & Percival R (2009) “Longitudinal benchmarking and alignment of a dynamic microsimulation model”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 20 pages, available from http://www.microsimulation.org/IMA/Ottawa_2009.htm

Kennedy S, McDonald JT & Biddle N (2007) “The healthy immigrant effect and immigrant selection: evidence from four countries”, 44 pages, downloaded from www.personal.buseco.monah.edu.au 18/3/10

Kippen R, Gray E & Evans A (2005) “The impact on Australian fertility of wanting one of each”, People and Place, vol 13 no 2 p13-20

Klevmarken A & Olovsson P (1993) “Direct and behavioural effects of income tax changes - simulations with the Swedish model MICROHUS”, in “Microsimulation and public policy”, ed Harding A, North-Holland, Amsterdam, 1996, p203-232

Klosgen W (1983) "Software implementation of microanalytic simulation models - state of the art and outlook", in "Microanalytic simulation models to support social and financial policy", eds Orcutt GH, Merz J & Quinke H, North-Holland, Amsterdam, 1986, p475-491

Krupp H (1983) "Potential and limitations of microsimulation models", in "Microanalytic simulation models to support social and financial policy", eds Orcutt GH, Merz J & Quinke H, North-Holland, Amsterdam 1986, p31-41

Laws P & Hilder L (2008) “Australia’s mothers and babies 2006”, AIHW National Perinatal Statistics Unit, Sydney, December, ix + 94 pages

Leblanc N, Morrison R & Redway H (2009) “A match made in silicon: Marriage matching algorithms for dynamic microsimulation”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 27 pages

Li J & O’Donoghue C (2011) “Evaluating alignment methods in dynamic microsimulation models”, paper presented to the third general conference of the International Microsimulation Association, Stockholm, June 8-10, 33 pages

Lymer S (2009) “APPSIM - modelling health”, National Centre for Social and Economic Modelling, Canberra, Working Paper no 13, March, 41 pages

Mark K and Karmel T (2010) “The likelihood of completing a VET qualification: A model-based approach”, NCVER, Adelaide

Mazzaferro C & Morciano N (2009) “Measuring intragenerational and intergenerational redistribution in the Italian social security system”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 30 pages

McKay S (2003) “Dynamic microsimulation at the US Social Security Administration”, presented at the International Microsimulation Conference on Population Ageing and Health, Canberra, 11 December, 7 pages

Metropolis N (1987) “The beginning of the Monte Carlo method”, Los Alamos Science (1987 special issue dedicated to Stanislaw Ulam) p125-130

Miller EJ (2009a) “Simulating cities: the ILUTE project”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 22 pages

Miller EJ (2009b) personal communication

Miller EJ, Chingcuanco F, Farooq B, Habib KMN, Habib MA & McElroy D (2008) “Development of an operational integrated urban model system - volume IV - demographic and labour market updating”, University of Toronto, December, 37 pages

Mitton S, Sutherland H & Weeks M (2000) “Microsimulation modelling for policy analysis - challenges and innovations”, Cambridge University Press, xxi + 331 pages

Morand E, Toulemon L, Pennec S, Baggio R & Billari F (2010) “Demographic modelling: the state of the art”, SustainCity Working paper 2.1a, Ined Paris, downloaded from www.sustaincity.org/Publications/WP2.1a 18/10/11

Morciano M (2009) personal communication, July 1

Morrison R (1998) “Overview of DYNACAN - a full-fledged Canadian actuarial stochastic model designed for the fiscal and policy analysis of social security schemes”, slightly adapted by Dussault B for inclusion on the website of the International Actuarial Association website, downloaded 9/10/09 from www.actuaries.org/CTTEES_SOCSEC/Documents/dynacan.pdf

Morrison R (2006) “Make it so: event alignment in dynamic microsimulation”, DYNACAN Team, 7th floor, Narono Building, 360 Laurier Avenue West, Ottawa, Ontario, May, 21 pages www.ssb.no/misi/papers/morrison-rick_makeitso-oslo-2.doc

Morrison R (2008) “Validation of longitudinal microsimulation models: DYNACAN practices and plans”, National Centre for Social and Economic Modelling, University of Canberra, Working Paper No 8, March, 41 pages

Muller GP (1983) "Data structure requirements of microanalytic simulation models" in "Microanalytic simulation models to support social and financial policy", eds Orcutt GH, Merz J & Quinke H, North-Holland, Amsterdam, 1986, p495-516 pages

National Statistics (2007) “Variations persist in life expectancy by social class”, London, News Release October 24, 8 pages (www.statistics.gov.uk)

NCVER (2010) “Historical time series of vocational education and training in Australia, from 1981”, downloaded 6/12/11 from www.ncver.edu.au

Nelissen JMH (1993) chapter 12 in “Microsimulation and public policy”, ed Harding A, North-Holland, Amsterdam

O’Donoghue C (2001) “Dynamic microsimulation: a methodological survey”, Brazilian Electronic Journal of Economics, 4(2), 77 pages, www.microsimulationa.org/IMA/BEJE/BEJE_4_2_2.pdf downloaded 16/11/09

O’Donoghue C, Leach RH & Hynes S (2007) “Simulating earnings in dynamic microsimulation models”, in “New frontiers in microsimulation modelling”, eds Zaidi A, Harding A & Williamson P (eds), European Centre Vienna, Ashgate, p381-412

O’Donoghue C, Lennon J & Hynes S (2009) “The life-cycle income analysis model (LIAM): A study of a flexible dynamic microsimulation modelling computing framework” (2009), International Journal of Microsimulation 2(1) p16-31

Orcutt GH (1957) "A new type of socio-economic system", Review of Economics and Statistics, 58, p773-797

Orcutt GH, Greenberger M, Korbel J & Rivlin A (1961) "Microanalysis of socioeconomic systems - a simulation study", Harper & Brothers, New York, xviii + 425 pages

Orcutt GH, Caldwell S & Wertheimer R (1976) “Policy exploration through microanalytic simulation”, The Urban Institute Washington DC xvi + 370 pages

Orcutt GH, Merz J & Quinke H (1986) "Microanalytic simulation models to support social and financial policy", North-Holland, Amsterdam, xvii + 553 pages

Payne A, Percival R & Harding A (2008) “APPSIM - modelling education”, National Centre for Social and Economic Modelling, Canberra, Working Paper no 11, June, v + 37 pages

Pennec S & Bacon B (2007) “APPSIM - modelling fertility and mortality”, National Centre for Social and Economic Modelling, Canberra, Working Paper no 7, December, v + 51 pages

Percival R (2007) “APPSIM - software selection and data structures”, NATSEM working paper no 3, April, 15 pages

Perese K (2002) “Mate matching for microsimulation models”, Technical Paper 2002-3, Long-term Modelling Group, Congressional Budget Office, Washington DC, 32 pages

Petrovic V (2009) “High-level design overview of DYNACAN, a Canadian dynamic longitudinal microsimulation model”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 20 pages

Pettersson T (2009) personal communications, June 24 & September 22

Pettersson T (2011) personal communications, October 21

PIneles-Mark C, Schwabish JA & Simpson M (2009) “Incorporating immigrant flows into microsimulation models”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 34 pages

Productivity Commission (2011) “Disability care and support”, Canberra, Draft report, two volumes

Real Estate Institute of Australia (2008), quarterly data on median house prices and net rental yields for 3 bedroom houses for each capital city, supplied by email February 2008

Reserve Bank of Australia (2011) “Table F5 Indicator lending rates”, downloaded from www. rba.gov.au on 25/9/11

Redway H (2009) personal communication, June 26

Roberts D (2010) “Transition probabilities from Mark-Karmel”, NCVER, spreadsheet supplied November 8

Sabelhaus J (2009) personal communication, November 17

Sarjeant HB (1998), program creating synthetic households from Australian census data for local statistical areas

Simpson (2011), Congressional Budget Office, personal communication, October 8

Simpson M & Topoleski J (2005) “Quantifying uncertainty in the analysis of long-term social security projections”, Congressional Budget Office, November, 40 pages, www.cbo.gov/publications

Spielauer M (2007) “Dynamic microsimulation of health care demand, health care finance and the economic impact of health behaviours: survey and review”, International Journal of Microsimulation 1(1) p35-53

STATA (2007) “STATA longitudinal/panel-data reference manual release 10”, Stata Press, Texas, iii + 474 pages

Statistics Canada (2009) “Microsimulation”, downloaded from http://www.statcan.gc.ca/microsimulation/index-eng.htm November 11

Strategic Forecasting (2002) “About Strategic Forecasting”, www.stratetic.forecasting.com, downloaded 23/6/09

Voas D & Williamson P (2000) “An evaluation of the combinatorial optimization approach to the creation of synthetic microdata”, International Journal of Population Geography 6, p349-366

Walker AE, Butler JRG & Colagiuri S (2008) “Cost-benefit model system of chronic diseases in Australia to assess and rank prevention and treatment options - proposed approach”, ACERH Research Report No 3, February, iv + 39 pages

Walker AE, Butler JRG & Colagiuri S (2009) “Economic model of chronic diseases in Australia: the prototype”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 14 pages

Watson L (2008) “HILDA user manual - release 6”, University of Melbourne, March 13, 165 pages

Wertheimer R, Zedlewski SR, Anderson J & Moore K (1983) "DYNASIM in comparison with other microsimulation models", in "Microanalytic simulation models to support social and financial policy", eds Orcutt GH, Merz J & Quinke H, North-Holland, Amsterdam 1986, p187-222

Wolfson MC (2009) “Preface - Orcutt’s vision 50 years on”, in “New frontiers in microsimulation modelling”, eds Zaidi A, Harding A & Williamson P, European Centre Vienna, p21-30

Wu B, Birkin MH & Rees PH (2009) “A hybrid spatial microsimulation model for decision support in demographic planning”, paper presented to the second general conference of the International Microsimulation Association, Ottawa, June 8-10, 17 pages, downloaded from http://www.microsimulation.org/IMA/Ottawa_2009.htm 15/11/09

Yevick D (2005) “A first course in computational physics and object-oriented programming with C++”, Cambridge University Press, xiii + 403 pages

GLOSSARY

ABS	Australian Bureau of Statistics
ACERH	Australian Centre for Economic Research on Health
ACFI	Aged Care Funding Instrument
ACT	Australian Capital Territory
AIHW	Australian Institute of Health and Welfare
ANU	Australian National University
APPSIM	NATSEM household microsimulation model for Australia
APRA	Australian Prudential Regulatory Authority
Calibration	Adjusting model assumptions to give reasonable projections
CAPP-DYN	Microsimulation model for Italy
CBOLT	Congressional Budget Office microsimulation model for USA
Closed model	A microsimulation model in which the only new persons are births or immigrants
CORSIM	Strategic Forecasting microsimulation model for USA
Creep	Systematic changes between repeated simulation runs, as a result of programming errors
Cross-sectional	A microsimulation model of all the persons alive at a particular time
CSV	Comma-separated variables
DEEWR	Department of Education, Employment and WorkPlace Relations
Destinie	Microsimulation model for France
Dynamic model	A microsimulation model that updates each attribute of each person in each time interval
CURF	Confidential unit record file released by ABS
CSIRO	Commonwealth Scientific & Industrial Research Organization
CSV	Comma separated values, a widely used file format
DYNACAN	Microsimulation model for Canada
DYNAMOD	NATSEM microsimulation model for Australia
DYNASIM	Urban Institute microsimulation model for USA
ERP	Estimated resident population (estimated annually by ABS)
Exit	At least one person leaving a household, with at least one person remaining
FOI	Freedom of Information
HILDA	Household, Income and Labour Dynamics in Australia Survey
HSF	Household Sample File (1% census sample released by ABS)
ILUTE	Integrated Land Use, Transportation, Environment model of Toronto
INAHSIM	Microsimulation model for Japan
IVF	In-vitro fertilisation
LIAM	Life-Cycle Income Analysis Model
LLR	Log-likelihood reduction
MIDAS-BE	Microsimulation model for Belgium, based on LIAM
MOSART	Microsimulation model for Norway
MSE	Mean standard error
MSR	Major statistical region, as used in HILDA
NATSEM	National Centre for Social and Economic Modelling, University of Canberra
NCVER	National Centre for Vocational Education Research
NSW	New South Wales
ODD	Order of decreasing difficulty
PENSIM2	Microsimulation model for UK
POLISIM	Microsimulation model for USA
PRISME	Microsimulation model for France
RBA	Reserve Bank of Australia

REIA	Real Estate Institute of Australia
SA	South Australia
SAGE	Microsimulation model for UK
SD	Standard deviation
SDAC	Survey of Disability, Ageing and Carers
SE	Standard error
Sfb 3	Sonderforschungsbereich 3, microsimulation model for the FDR
SESIM	Microsimulation model for Sweden
SIHC	Survey of Income and Housing Costs, Australia
SMILE	Microsimulation model for Ireland, based on LIAM
Superannuation	Australian superannuation funds intended to provide retirement benefits separately from social security entitlements
STATA	A statistical package created by Stata Corporation
SVERIGE	Microsimulation model for Sweden
TAFE	Technical and further education
TFR	Total fertility rate – the sum of the age-specific fertility rates
WA	Western Australia