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

© 14 December 2011 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 Jeromney 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.


- 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 of 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
   1.1 Scope of this thesis .................................................................................................. 1
   1.2 Orcutt proposals for microsimulation, and their implementation ...................... 1
   1.3 Limited range of uses of present models ........................................................... 1
   1.4 Using sampling with loaded probabilities to get shorter simulation intervals .... 2
   1.5 Ways to reduce run times .................................................................................... 2
   1.6 Ways to increase spatial disaggregation ............................................................. 2
   1.7 Ways to reduce distortions from event alignment and matching ....................... 2
   1.8 Programming languages and data structures ..................................................... 3
   1.9 Ways to reduce errors .......................................................................................... 3
   1.10 Data needs and sources ..................................................................................... 3
   1.11 Simulating demographic events and household changes .................................. 3
   1.12 Simulating education, occupation, employment and earnings .......................... 4
   1.13 Simulating superannuation contributions and fund balances ............................ 4
   1.14 Simulating housing ............................................................................................ 4
   1.15 Simulating disease, disability and need for residential care ............................. 4
   1.16 Meets the needs of potential users ..................................................................... 4
   1.17 Assistance to other modellers .......................................................................... 5

2. LITERATURE REVIEW....................................................................................................... 6
   2.1 Definitions of microsimulation ........................................................................... 6
   2.2 Orcutt’s 1957 proposals for household and firm simulations ............................. 6
   2.3 Some national cross-sectional dynamic models of households ....................... 7
   2.4 Cross-sectional, closed, dynamic, discrete-time models .................................... 8
   2.5 Links between national microsimulation models ............................................ 9
   2.6 High costs of national microsimulation models ............................................. 10
   2.7 Uses for national models ................................................................................... 10
   2.8 Demise of some national microsimulation models ........................................ 12
   2.9 First documented uses of key household modelling techniques .................... 13
   2.10 Immediate versus batch matching .................................................................. 13
   2.11 Event alignment ............................................................................................... 13
   2.12 Error detection and model validation ............................................................. 14
   2.13 Computer techniques ...................................................................................... 14
   2.14 Technical details and run times for national models ....................................... 15
   2.15 Reasons for differences between run times ................................................... 16
   2.16 Advantages of short run times ....................................................................... 17
   2.17 Advantages of short simulation cycles .......................................................... 17
   2.18 Advantages of fine geographic subdivisions ................................................. 18
   2.19 Data synthesis for small areas ........................................................................ 19

3. SIMULATION BY SAMPLING WITH LOADED PROBABILITIES .................................. 21
   3.1 All-case simulations ............................................................................................ 21
   3.2 Simulation by sampling with loaded probabilities .......................................... 21
   3.3 A trivial case: uniform event probabilities ...................................................... 22
   3.4 Non-uniform event probabilities, with no losses from pool ............................ 22
   3.5 Non-uniform event probabilities, with losses from pool ................................. 22
   3.6 Numbers of draws when using sampling with loaded probabilities ............... 24
   3.7 Tests on accuracy ............................................................................................. 24
   3.8 Example of the use of pools in simulating deaths ........................................... 25
   3.9 Pools used to test sampling with loaded probabilities .................................... 25
   3.10 Run times to simulate 175,000 Australian for a year ...................................... 26
   3.11 Run times for projections for up to 50 years ............................................... 27
   3.12 Checks on projected event numbers .............................................................. 27
   3.13 One-year projection results with each event simulated separately ............... 28
   3.14 One-year projection results with all events simulated together .................... 29
   3.15 Comparisons between normal and reverse order simulations ..................... 29
   3.16 Comparisons between all-case and loaded 50-year projections .................... 30
   3.17 Checks on event number standard deviations .............................................. 30
   3.18 Coefficients of variation for one-year projections ........................................ 31
   3.19 Computational aspects .................................................................................... 31
   3.20 Likely applications .......................................................................................... 32
   3.21 Conclusions ..................................................................................................... 32

4. EVENT AND STATE ALIGNMENT .................................................................................. 34
   4.1 Alignment needs for household microsimulations ......................................... 34
   4.2 Consistency with beliefs about the future ...................................................... 34
   4.3 Use of alignment to eliminate stochastic variation ........................................ 35
   4.4 Obtaining reasonable projections without alignment ....................................... 35