

Using Microsimulation to Create Synthetic Small-Area Estimates from Australia's 2001 Census

Paul Williams

Background

The aim of the project is to greatly improve the decision-support tools available to State and Territory governments by providing them with:

- far more detailed small area data than has previously been available, via the creation of a synthetic small area database, plus
- the capacity to assess the current and future impact of possible policy reforms and of likely demographic, social and economic changes at the small area level, through the construction of microsimulation models on top of the synthetic small area database.

In the past, all policy makers have been limited by a lack of detailed small area data. The key source of small area data in Australia is the five yearly Census of Population and Housing conducted by the ABS. The crucial advantage of the Population Census is that it contains *detailed small area* socio-demographic information. At the greatest level of geographic disaggregation, results are available for each of about 37,000 Collection Districts in Australia (with a Collection District containing about 200 households).

However, there are two important limitations to the Census data from the perspective of policy makers. First, the amount of information collected from each household is relatively limited. For example, only gross household income is collected and then only in broad ranges of income. There is also no information about the social security receipt, income sources, wealth or expenditure of households.

Second, unlike many other ABS collections, the full Census results are not publicly available as a unit record file. (A unit record file would contain the responses of each household as a separate record in a database.) Instead, output for the whole Census file is only available as a pre-defined series of tables for each Collection District (which commonly contain many confidentialised cells which limits their usefulness). This means, for example, that relationships between characteristics of interest cannot be fully explored (such as age by income by educational qualifications). In addition, microsimulation models – which have by now become of great importance to national policy makers across the industrialised world – cannot be constructed on top of the pre-defined tables.

This project aims to fill this major gap in the information available to State and Territory governments by creating synthetic small area data for each Census Collection District. This is achieved by matching the characteristics of the households from other datasets with aggregated information for each of the Census Collection Districts. Results are then available at the Collection District level, which can be used for further analysis. Results at a level below the Collection District are not generated, so that privacy concerns are satisfied.

Microsimulation

During the past two decades microsimulation models have revolutionised the quality of information about the distributional and revenue impacts of policy changes available to policy makers in industrialised countries. Microsimulation models begin with a dataset that contains detailed information about the characteristics of each household and person within a sample survey or an administrative database. National level results are then calculated by adding together the results for each individual household. Microsimulation models have been assessed by the OECD (1996) and the US National Academy of Sciences (Citro and Hanushek, 1991) as one of the most useful tools available to policy makers for assessing the distributional consequences of possible policy changes. Such models are now widely used by national policy makers throughout the developed world (Harding, 1996; O'Donoghue and Sutherland 1996; Gupta and Kapur, 2000).

However, an important limitation of the models to date has been that results have only been available at the *national* level or, at best, at a State level. This is because the existing models have been constructed on top of ABS sample survey data, in which most geographic detail has been suppressed by the ABS to protect the confidentiality of those households who took part in the survey. Thus, it has not been possible in the past using these models to predict the *spatial* impact of possible policy changes upon the household sector.

To overcome this problem, during the past two years the National Centre for Social and Economic Modelling (NATSEM) located at the University of Canberra has begun the construction of spatial microsimulation models. To date, these new types of model have combined data from the Population Census and the ABS sample surveys (such as the Household Expenditure Surveys). The new spatial microsimulation modelling techniques developed at NATSEM blend the Census and sample survey data together to create a synthetic unit record file of households for every Collection District. The first model to be constructed by NATSEM using these new techniques was the Marketinfo/99 model, which provided detailed regional expenditure and income estimates (Harding et al, 1999). This model has since been used in the analysis of poverty by statistical sub-division in the ACT (Harding et al, 2000) and for the analysis of poverty levels by postcode in Australia (Lloyd et al, 2001). The data have also been utilised by a range of major Australian organisations, including Lend Lease, the Commonwealth Bank and KPMG, for determining where to locate shopping centres and for examining the expenditure patterns of target customers.

The new modelling is also being used as the basis of a long-term major strategic planning model for provision of social security payments, which involves simulating the access patterns and characteristics of social security customers by small area. In addition, NATSEM has won a grant from the Australian Housing and Urban Research Institute to extend the existing modelling to simulate the Commonwealth rent assistance scheme and the impact of possible changes to that scheme (such as payment rates that vary with location).

These efforts to develop spatial microsimulation are at the leading edge internationally. Other countries where spatial microsimulation is being developed comprise the highly regarded CORSIM project in the US (Caldwell et al, 1998); and Leeds and Liverpool Universities in the UK (Voas and Williamson, 2000, Ballas and Clarke, 1999).

The initial investment has created a prototype model and demonstrated the feasibility of constructing spatial microsimulation models. This project is expected to:

- refine the techniques used to create the synthetic small area data.
- provide much more extensive validation of the outcomes.
- add additional characteristics to the simulated households.
- simulate the impact of tax, social security and other changes at a spatial level.
- develop techniques for ageing the small area data forward through time.
- initiate linkages between NATSEM's dynamic modelling of wealth and superannuation and the spatial projections of households.

The project seeks to create robust and validated spatial datasets and models that State and Territory policy makers can have confidence in and use. The spatial database and modelling framework will be of benefit to all levels of government in Australia – as well as a wide range of other organisations interested in the spatial distribution of socio-economic characteristics and the spatial consequences of policy and other changes.

Approach and timeline

The new regional microsimulation modelling techniques developed at NATSEM during the past few years blend the Census and ABS sample survey data together to create a synthetic unit record file for every Collection District. To date, NATSEM's efforts have focussed upon the ABS Household Expenditure Survey (HES), although efforts are currently underway to extend the methodology so as to 'regionalise' other sample survey data. The existing model first recodes the HES and Census variables to be comparable, and then reweights the HES, utilising detailed sociodemographic profiles from the Census Basic Community Profiles. This is done for each Collection District separately, and a reweighted HES unit record file is generated for each District. Results and output are generated at the Collection District level or higher levels of geographic aggregation. Results are not available at a level below the Collection District.

2003

The new 2001 Census Basic Community Profiles will be released late in 2002. The first task envisaged for this project for 2003 is to create a new spatial microsimulation database, 'regionalising' the 1998-99 ABS Household Expenditure Survey by creating synthetic households whose characteristics match those shown in the new 2001 Census data. At the completion of this phase there will be a synthetic database of households for each Collection District in 2001, with far more detail than is available in the Census about their estimated income, income sources, spending patterns and housing costs. At the conclusion of this phase, a database will be produced containing about 50 variables for each Collection District within their State or Territory. The 50 variables might, for example, include average gross income for each CD, average

investment income, average mortgages paid by labour force status of household head, average rent paid, poverty rates (using a range of poverty measures), housing affordability measures and relevant expenditure measures (such as expenditure on food as a percentage of total expenditure).

This new database will immediately answer many questions about the income of households and about possible areas of unmet need and poverty at the small area level. Analysis can also be undertaken by gender to examine, for example, whether there are particularly high poverty rates or low income levels for women or for men living in different areas. Similarly, questions of housing affordability and access are very important policy issues.

A second key task, for the second half of 2003, is to add estimated assets to the records of each of the synthetic households. Respondents to the Household Expenditure Survey are asked details about their income from various investments, such as dividends from shares and rental income from investment properties. In recent years NATSEM has developed income capitalisation techniques, to estimate the value of the assets that underlie the generation of these investment income flows (Kelly, 2001; Harding et al, 2002a). These techniques will be used to estimate the value of the superannuation, own business, shares, cash holdings and investment properties owned by each of the synthetic households. At this stage the value of owner-occupied housing will not be estimated, as there are major difficulties in accurately estimating the current value of housing by Collection District and household type. However, we hope in future work to draw upon sources of data available to the State and Territory to estimate this, and have also established links with Baycorp Advantage Pty Ltd, a company that has invested millions of dollars in a model to estimate the likely sales price of houses, with a current coverage of about 2/3 of Australian households.

In the second half of 2003 the work will also concentrate upon constructing microsimulation models on top of the newly created synthetic microdatabase. The existing program rules for income tax, the Medicare levy, and social security and family payments will be replicated in computer code so that, for example, the amount of income tax paid and age pension received under the current program rules can be estimated. The amount of Goods and Services Tax (GST) paid under the existing rules will also be estimated, drawing upon recent simulation in this area by NATSEM (Harding et al, 2002b).

The simulation of the rules of government programs is extremely complex, but this component of the work will draw upon the thousands of lines of computer code already written by NATSEM to undertake this task at the national level. However, simulating social security receipt and tax payments at the spatial level will require additional work and the development of new techniques. This is because the results arising from the initial programming for the simulated households will have to be benchmarked against other data, such as taxation statistics by postcode and Centrelink client numbers by region. In addition, it will be important to comprehensively check that when the results for the synthetic households are added together, they do sum to comparable national and State benchmark data. For example, it is possible that estimated GST collections by State produced by summing the estimated GST paid by the synthetic households living in each state may not match the estimates contained within government finance statistics. The accurate simulation of social security

payments and tax liabilities at the spatial level is thus considerably more complex than the simulation at the national level (which NATSEM has been doing for the past eight years).

At the conclusion of this phase an enhanced database will be produced for each Collection District summarising the estimated characteristics of households living in each particular District. This will contain all of the average values for each Collection District mentioned earlier, but now expanded to include such variables as average total wealth for households living in that Collection District, average accumulated superannuation, average income tax paid, average GST paid, average social security received, average family payments received, estimated percentage of households with an investment property, percentage of households with less than \$100,000 in assets, percentage of households that are self-funded retirees or sole parents on pension, and so on.

Users of the database will thus have the capacity to undertake comprehensive analysis of the *spatial patterns* of income, wealth and spending, along with estimated social security payments and income tax and GST incidence. As noted earlier, questions of housing affordability and access are particularly important. After the simulation of income tax, it will be possible to calculate the disposable (after-income-tax) income of the synthetic households, and develop new housing affordability measures (such as the percentage of after-tax income devoted to housing). A range of other housing affordability variables will also be developed and added to the database. It is envisaged that, by the completion of this phase, a database containing about 100 variables for every Collection District within their State or Territory will be developed.

2004

One of the initial key tasks for 2004 will be to develop the capacity to *change* the rules of taxation, social security, housing assistance, and other programs and predict the consequent spatial distributional effects. For example, it would be possible to simulate the spatial distributional impact of changes in rent assistance, reductions in the income tax threshold, an increase in the top marginal tax rate, an increase in the GST rate from 10 to 11 per cent, or a liberalisation in the social security income test rules.

Having established the initial version of the new spatial databases and models, a significant part of 2004 will be devoted to determining whether the accuracy of the modelling can be improved. Critical to this phase will be collaboration with Chief Investigator Williamson, who will again visit NATSEM for a few weeks during this year and who has recently completed a 2 ½ year UK research council funded programme on the creation of spatial synthetic microdata (Huang and Williamson, 2001). Also critical will be the assistance of the ABS. There are four areas where further work on techniques is expected to be particularly valuable, namely assessment and development of:

- improved methods of optimising the household weights;
- methods of uprating the HES data and, conceivably, the Census data, to reflect changes since the data were collected;

- new methods of validation;
- quality assurance techniques for spatial microdata; and
- techniques to describe the robustness of the results for each CD (e.g. an index that summarises whether the results appear particularly good or bad for a selected CD).

One issue, for example, is that currently NATSEM creates the synthetic households by weighting to the publicly available Census Basic Community Profiles. Additional detail is available in the original Census data and it is possible that, by weighting to a more complex matrix, the characteristics of the synthetic households might be able to be made to match the characteristics of real households more closely.

Similarly, considerable work will be devoted to validation and the systematic testing of the reliability of the results for different spatial units. For example, is there a trade off between accuracy and the size of the spatial unit for which synthetic households are being created? Is it the case that the results are particularly good for, say, 80 per cent of Collection Districts, but problematic for the remaining 20 per cent? Are results particularly good for metropolitan areas but less reliable for remotely settled areas?

This phase of the work will also draw upon the accumulated regional expertise and data of key users of small area data. They will be able to provide feedback about which CD results appear unusual to them, given their knowledge of that particular region. This 'on-the-ground' expertise is expected to be invaluable in further validating the model. Overall, the results from this phase will be used to refine the modelling and to increase our understanding of and confidence in the synthetic data.

2005

A key task for 2005 is to develop and test techniques for ageing the spatial microdata forward by 5 to 20 years. 'Static ageing' techniques are well developed in microsimulation (Harding, 1993) and are regularly used by NATSEM when, for example, ageing out-of-date ABS survey data up to the current world. Such static ageing techniques traditionally involve uprating (or inflating) incomes and housing costs to current or projected future levels and reweighting data to account for expected demographic or labour force changes (such as an ageing population or a reduction in unemployment).

The likely income and asset position of the ageing baby boomers is a key issue when planning future residential development and infrastructure. Because the economic capacity of the ageing baby boomers in their retirement will emerge as such an important issue, the goal is to inform the static ageing techniques with some results from NATSEM's DYNAMOD dynamic microsimulation model (Kelly et al, 2001; King et al, 1999a, 1999b). Thus given the projections of the likely age/gender population structure of particular small areas in, say, 20 years, NATSEM could then change the weights attached to households in the synthetic small area database to reflect those new population targets. This would give an initial and immediate impression of the likely characteristics of those expected to be living in these areas in 20 years time. This would then be complemented by some imputation of the likely

wealth of these households, drawing upon existing NATSEM modelling of wealth and superannuation in the long term (Kelly et al, 2001).

The other key goal for 2005 is to prepare a series of technical papers documenting the construction of the GeoSim databases and models. Although documentation is a time-consuming and thus expensive task, NATSEM's experience has shown that it is an essential part of the process of keeping models alive in the long term.

The final goal is to use the new modelling capacity to analyse a range of questions of interest, with the results to be presented at conferences and then written up for journal articles.

Australian Bureau of Statistics involvement

State and Territory governments have in the past been limited by the lack of detailed data on the socio-economic characteristics of households at the small area level. Given the lack of data available in the publicly available CDATE tables, such governments have often commissioned the ABS to produce specialised data or analysis from the Census unit record file (a data source which cannot be accessed by anyone outside the ABS). However, the data from the Census is limited, compared with the data available in other ABS national sample surveys - and detailed cross-classified tables at Collection District level typically contain many confidentialised cells that limit their usefulness.

Data limitations have been particularly pronounced for the smaller States and Territories seeking to understand more about the characteristics of their residents. For example, of the 7000 households included in the sample for the 1998-99 ABS Household Expenditure Survey, only 275 were from the Australian Capital Territory (ACT). As a result, even if the ACT government commissioned the ABS to provide further details about these households, the sample size is so small that little disaggregation can be undertaken.

Recognition of these existing data limitations is one of the important reasons lying behind the ABS's proposed participation in this project. While the ABS cannot grant non-ABS officers access to the original Census unit record data, the ABS is prepared to commit resources to providing substantial output from the Census and other data as part of the work involved in creating and improving synthetic small area data. The ABS can see major advantages in detailed small area synthetic data being widely available to government and other agencies. In particular, imputation of detailed income and wealth data onto synthetic small area household data is one way to enhance the usefulness of the Census, in which very detailed income and wealth data will never be collected.

One important aspect of the ABS involvement is that the robustness of the simulation of additional characteristics onto the records of households can be checked. For example, while NATSEM and other non-ABS users do not have access to information about the geographic location of respondents to the ABS Household Expenditure Survey, the ABS does have this information. Thus, the ABS is in a position to provide

advice about how reliable the imputation of income and spending data to the synthetic households has been at a regional level.

National Benefit

Regional issues have recently assumed much greater importance in Australia. There is a growing realisation that the gains from economic growth have not been equally distributed amongst different regions in Australia (Gregory and Hunter, 1995, Vinson, 2000).

While national policy makers have utilised microsimulation models for the past decade in Australia to assess the revenue and distributional consequences of policy change, it has not previously been possible to assess the consequences of policy change upon households at a spatial or small area level. The Census data is the only comprehensive data source in Australia with detailed information about the characteristics of households at the small area level. However, only a relatively limited range of information is collected about households in the Census and the full Census records of all households are not available for public use (with such household level records being required for the construction of microsimulation models). As a result, the quality of information about the socio-economic characteristics of households and about the spatial consequences of policy change available to State and Territory policy makers has been very limited.

The significance of this research proposal is thus that it will lead to very substantial — and greatly needed — improvements in the quality of information available to policy makers and researchers about the regional distributional impact of changes in tax, social security and other policies.

The project will involve the development of robust synthetic small area data and models that will be of immense use to all levels of government as well as other organisations. While the project is at the cutting edge and thus innately high risk, the potential benefits are enormous. There are hundreds of organisations within Australia that would benefit from detailed data about the socio-economic characteristics of households at a small area level. Such organisations range from governments engaged in needs-based planning for health and other community services, to private sector companies interested in assessing the incomes and expenditures of households living in different areas, to Federal policy makers evaluating the spatial impact of national changes in government policies.

This project is also likely to be of great benefit to regional and rural communities. For the first time, detailed information will be available about how their incomes and wealth compare to those of residents of metropolitan areas. The new spatial database and model will help both State and Territory governments in identifying areas of unmet need and in their planning of the spatial provision of government services.

The output from the research will be published in a series of conference and discussion papers and then in journal articles. NATSEM makes strenuous efforts to make its work accessible to non-economists and the public, including putting all papers on the NATSEM website, regularly presenting new research at conferences and contributing to the media. That NATSEM is successful in communicating its

results is evidenced by the intense public interest in NATSEM's research, which regularly receives extensive national media coverage. The Centre's website received around 1.5 million hits last year and about 200,000 copies of the Centre's publications were viewed and downloaded in 2001

References

- Ballas, D and G Clarke, 1999, 'Modelling the local impacts of national social policies: A microsimulation approach', Papers presented at the 11th European Colloquium on Theoretical and Quantitative Geography, Durham Castle, Durham, England, 3rd to 7th September.
- Caldwell, S. B., Clarke, G. P. and L.A. Keister, 1998, Modelling Regional Changes in US Household Income and Wealth: A Research Agenda, *Environment and Planning C: Government and Policy*, Vol 16, pp 707-722.
- Citro, C. F. and E. A. Hanushek, 1991, *The Uses of Microsimulation Modelling, Vol 1: Review and Recommendations*, National Academy Press, Washington
- Gregory, R. G. and Hunter, B., (1995), 'The macro economy and the growth of ghettos and urban poverty in Australia', Discussion Paper No 325, (paper presented as the Telecom Address at the National Press Club, April 22), Centre for Economic Policy Research, Australian National University.
- Gupta, A and V Kapur (eds), *Microsimulation in Government Policy and Forecasting*, Contributions to Economic Analysis Series, North Holland, Amsterdam.
- Harding, A., 1993, *Lifetime Income Distribution and Redistribution: Applications of a Microsimulation Model*, Contributions to Economic Analysis Series, Amsterdam, North Holland.
- Harding, A., *Microsimulation and Public Policy*, (ed), Contributions to Economic Analysis Series, Amsterdam, North Holland, 1996.
- Harding, A., Hellwig, O., Bremner, K. and Robinson, M., 'Geodemographics of the Aged: Where they live, What they buy'" Paper presented at the 'Geodemographics of Ageing in Australia' Symposium, Brisbane, 2 December, 1999.
- Harding, A., Lloyd, R., Hellwig, O and Bailey, B. 2000, *Building the Profile: Report of the Population Research Phase of the ACT Poverty Project*, Taskgroup Paper No 3, ACT Poverty Task Group, ACT Government, Canberra.
- Harding, A., King, A and Kelly, S., 2002a, Incomes and Assets of Older Australians: Trends and Policy Implications, *Agenda*, Volume 9, Number 1.
- Harding, A., Warren, N., Beer, G., Phillips, B. and Osei, K., 2002b, The Distributional Impact of Selected Commonwealth Outlays and Taxes, paper prepared for the Review of Commonwealth-State Funding, 14 March 2002.
- Huang Z and Williamson P (2001) *A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata*. Working Paper 2001/2, Population Microdata Unit, Department of Geography, University of Liverpool, Liverpool L69 3BX. <http://pcwww.liv.ac.uk/~william/microdata>
- Kelly, S, 2001, 'Trends in Australian Wealth – New Estimates for the 1990s', 30th Annual Conference of Economists, University of Western Australia, 26 September.
- Lloyd, R., Harding, A. and Greenwell, H, 2001, 'Worlds Apart: Postcodes with the Highest and Lowest Poverty Rates in Today's Australia' Paper prepared for the National Social Policy Conference 2001. Sydney, Australia. July (about to be published by SPRC).
- Kelly, S., Percival, R. and Harding, A. 2001, 'Women and Superannuation in the 21st Century: Poverty or Plenty?' Paper presented to SPRC National Social Policy Conference 2001, University of NSW, July*
- King, A., Bakgaard, H. and Robinson M., 1999a, 'DYNAMOD-2: An Overview', Technical Paper No. 19, NATSEM, University of Canberra, December.

- King, A., Bakgaard, H. and Robinson M., 1999b, 'The Base Data for DYNAMOD-2', Technical Paper No. 20, NATSEM, University of Canberra, December.
- O'Donoghue, C and Holly Sutherland (eds), 2000, *Microsimulation Modelling for Policy Analysis: Challenges and Innovations*, Cambridge University Press, Cambridge.
- OECD (Organisation for Economic Cooperation and Development), 1996, *Policy Implications of Ageing Populations: Introduction and Overview*, OECD Working Paper no. 33, Paris.
- Treasurer (1998), Document No 2, Documents Related to Household Expenditure Survey tabled in Parliament, 11 November.
- Voas, D and Williamson, P, 2000, 'An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata', *International Journal of Population Geography*, vol 6, pp 348-366.
- Vinson, Tony, 2000, *Unequal in Life: The Distribution of Social Disadvantage in Victoria and NSW*, The Ignatius Centre, Melbourne, August.
- Warren, N., Harding, A., Robinson, M., Lambert, S. and Beer, G. 'Distributional Impact of Possible Tax Reform Packages', *Main Report*, Senate Select Committee on a New Tax System, Senate Printing Unit, April, pp 445-508, 1999.