NATSEM
UNIVERSITY OF CANBERRA

Spatial Microsimulation: Preparation of Sample Survey and Census Data for SpatialMSM/08 and SpatialMSM/09

Technical Paper No.36

PREPARED BY

Rebecca Cassells, Ann Harding, Riyana Miranti, Robert Tanton and Justine McNamara

APRIL 2010

ABOUT NATSEM

The National Centre for Social and Economic Modelling was established on 1 January 1993, and supports its activities through research grants, commissioned research and longer term contracts for model maintenance and development.

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

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

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

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

ISSN 1443-5098

ISBN 978-1-74088-316-0

National Centre for Social and Economic Modelling

University of Canberra ACT 2601 Australia

170 Haydon Drive Bruce ACT 2617

Phone + 61 2 6201 2780

Fax + 61 2 6201 2751

Email natsem@natsem.canberra.edu.au

Website www.natsem.canberra.edu.au

Contents

ABSTRACT

Strong demand for small area estimates of the characteristics of households and for spatially detailed household microdata has prompted the development of spatial microsimulation techniques during the past decade. This Technical Paper first provides an overview of earlier approaches to 'synthetic estimation' and then describes some of the more recent advances in spatial microsimulation. An overview checklist of key issues that should be considered when preparing a sample survey for reweighting to Census (or other) small area benchmarks is also provided. The first steps in the creation of NATSEM's latest spatial microsimulation models – SpatialMSM/08 and SpatialMSM/09 - and the processes undertaken to prepare sample survey data for matching to 2006 Census benchmarks are described. Two sample survey unit record files were blended together and a range of processes were followed to allow these sample surveys to be reweighted to benchmarks for small areas calculated from the 2006 Census. Such processes included making variable definitions comparable between the surveys and the Census and imputing children, non-classifiable persons and individuals living in non-private dwellings onto the sample survey base files.

AUTHOR NOTE

Rebecca Cassells is a Senior Research Fellow at the National Centre for Social and Economic Modelling (NATSEM). Ann Harding is Professor of Applied Economics and Social Policy at NATSEM, Robert Tanton is Research Director, Riyana Miranti is a Research Fellow and Justine McNamara an Acting Principal Research Fellow, all at NATSEM.

AN APPROPRIATE CITATION FOR THIS REPORT IS

Cassells, R., Harding, A., Tanton, R., Miranti, R. and McNamara, J., (2010), 'Spatial Microsimulation: Preparation of Sample Survey and Census Data for SpatialMSM/08 and SpatialMSM/09', NATSEM Technical Paper No. 36

ACKNOWLEDGEMENTS

This project, 'Regional Dimensions: The Spatial Implications of Population Ageing and Needs-based Planning of Government Services', was funded by the Australian Research Council (Project No LP0775396), the NSW Department of Community Services, the Queensland Department of Premier and Cabinet, the Office of Economic and Statistical Research within the Queensland Treasury, the ACT Chief Minister's Department, the Victorian Department of Planning and Community Development, the Victorian Department of Education and Early Childhood, the Australian Bureau of Statistics, and NATSEM. NATSEM would like to thank these research partners for their enthusiasm and their on-going support to the project.

The authors would also like to gratefully acknowledge the on-going technical support and advice provided by our international chief investigator – Dr Paul Williamson from the University of Liverpool (UK). In addition, many past and present NATSEM staff have also made invaluable contributions to the Regional Dimensions projects and to NATSEM's work on developing spatial microsimulation.

This project was also co-funded by a Discovery Grant from the Australian Research Council (DP664429: Opportunity and Disadvantage: Differences in Wellbeing Among Australia's Adults and Children at a Small Area Level). NATSEM Chief Investigators on this grant are Professor Ann Harding and Robert Tanton. The authors would like to thank other Chief Investigators, Professor Fiona Stanley, Professor Bob Stimson, Professor Hal Kendig, Dr Sharon Goldfeld and the Australian Bureau of Statistics for their input to the broader work being undertaken through this grant.

GENERAL CAVEAT

NATSEM research findings are generally based on estimated characteristics of the population. Such estimates are usually derived from the application of microsimulation modelling techniques to microdata based on sample surveys.

These estimates may be different from the actual characteristics of the population because of sampling and nonsampling errors in the microdata and because of the assumptions underlying the modelling techniques.

The microdata do not contain any information that enables identification of the individuals or families to which they refer.

1 INTRODUCTION

1.1 SYNTHESISING SMALL AREA HOUSEHOLD DATA

There is strong demand for small area information about the characteristics of individuals and households and the small area impact of possible policy changes. First, such information is required, for example, by those government and non-government agencies with responsibility for allocating scarce resources to where they are most needed – ranging from the most effective placement of child care or aged care services to disability programs and services targeted towards youth-at-risk. Second, government often needs accurate information about the degree to which deprivation or disadvantage is concentrated in particular places, to inform social policy formation more generally. For example, if the income poverty rate within a society is 10 per cent of the population, then the most appropriate government response will be profoundly influenced by whether every suburb in the nation has a 10 per cent poverty rate or whether some suburbs are entirely populated by poor people and others contain no poor people at all. Third, an ability to estimate the spatial impact of a policy before the policy change is introduced helps to prevent the emergence of unintended small area consequences and reduce the risk associated with a policy change.

Yet, despite the compelling need for such small area data, it can often be very hard to obtain. National Censuses are typically conducted relatively infrequently and their extensive geographic detail comes at the price of containing only a limited range of information about households. Administrative data are sometimes geo-coded to provide a comparable level of geographic detail, but often only contain information essential to the provision of the services and usually lack socio-economic descriptors. National sample surveys, now typically available in unit record format to researchers, typically contain much richer information about a particular topic than the Census (income, health, expenditure) but usually suppress the geographic detail of respondents to protect privacy — and, even if that were not the case, usually provide too small a sample size to produce accurate small area estimates[1].

Not surprisingly, this data deficiency has led researchers to attempt to create synthetic small area estimates of household and other characteristics[2]. The techniques used to achieve this goal vary, but often involve fitting regression equations against sample survey

1 Unit record data (alternatively termed 'microdata') usually consist of thousands of individual records of persons, families or households in a computer readable format. Such microdata are the essential building block for microsimulation models, which in the past two decades have revolutionised the quality of information available to policy makers about the likely distributional impact of policy reforms that they are contemplating (Harding and Gupta, 2007a).

2 This data deficiency does not apply in the Scandinavian countries, which have an entirely different attitude to the US, Canada, and the rest of Europe and Australasia toward data on individuals. However, even though countries such as Sweden do not face the task of creating synthetic spatial microdata, they do still have to meet the challenge of constructing spatial microsimulation models on top of their real spatial administrative registers (Swan, 2000).

data that contains information about, say, health status, and then running the coefficients produced by those equations through small area Census data that contains the same predictive variables as contained in the sample survey. Thus, the dependent variable 'health status' may be estimated as a function of age, gender and occupation from the sample survey and then be imputed for each small area, using the same three explanatory variables within the Census data, resulting in a synthetic estimate of health status for each small area.

In theory, if the Census (or administrative) data were available at a unit record level, then this predictive equation could potentially be applied to every unit record for every individual within a particular small area and the results then aggregated for sub-groups within the small area population to provide an estimate of the proportion of individuals with, say, poor health, within a particular sub-group and small area. More often, however, researchers do not have access to unit record Census data for the entire country and thus have to apply their equations against published Census tables about the broad characteristics of the individuals living within each small area (Elazar and Conn 2005; Rao 2003; Pfeffermann 2002; and see EURAREA project for the EU – www.statistics.gov.uk/eurarea). Such 'synthetic estimation' or 'small area estimation' techniques have generally come from within the discipline of statistics.

Another possible approach, which has gained currency within the past decade, is the creation of synthetic spatial microdata. Some of the early research in this field was undertaken by geographers and concentrated upon whether it was possible to create small area specific microdata from the UK Census one per cent sample (Voas and Williamson 2000; Williamson et al. 1998). While various approaches to reconstructing spatially detailed microdata have been trialled, including data fusion and synthetic reconstruction (Voas and Wiliamson 2000, p. 349), the more successful endeavours essentially involve methods of reweighting the original sample survey data to match small area population targets shown in the relevant Census. To give a simplistic example, if the population Census indicates that one-half of all of the households within a particular small area are outright home owners, then one-half of all of the households selected from the national survey to 'populate' this particular small area will be outright home owners.

In practice, of course, researchers attempt to match the characteristics of households contained in the national survey not just to one piece of information about each particular small area (e.g. outright home owners) but to multiple cross-tabulations of household characteristics derived from the Census for each small area (such as age, gender, occupation, labour force status, socio-economic group and so on). Thus, as Ballas et al. (2006a, p.65) explain, these techniques 'involve the merging of Census and survey data to simulate a population of individuals within households (for different geographic units), whose characteristics are as close to the real population as it is possible to estimate'.

Once synthetic household microdata have been created for each small area, then it becomes feasible to enter the realms of microsimulation modelling. Such models were initially developed within the discipline of economics (Orcutt et al. 1986) and have today become very widely used by governments across the developed world for analysis of the fine-

grained distributional impact of possible changes in government programs (Harding and Gupta 2007b; Gupta and Kapur 2000; Mitton et al. 2000; Harding 1996). While most such models still assess the likely winners and losers from reforms in income tax and cash transfer programs (Harding et al. 2009a; Lloyd 2007; Immervoll et al. 2006), the past decade has also seen the rapid expansion of microsimulation models into other subject areas, such as health, housing and aged care (Gupta and Harding 2007; Brown et al. 2004; Harding et al. 2004). However, importantly, the overwhelming majority of these microsimulation models have been national models, constructed on top of national sample survey microdata and predicting the distributional impact of policy change for an entire country, rather than for a small region within a country.

A new development during the past decade has been the construction of spatial microsimulation models, constructed on top of the synthetic spatial microdata bases described earlier. This rapidly growing field now includes simulation of the small area impact of changes in income taxes, family payments and social security (Harding et al. 2009b; Chin et al. 2005); development of small area measures of poverty and housing stress (Tanton et al. 2009a; Tanton et al. 2009b; McNamara et al. 2007); small area modelling of Activities of Daily Living Status and need for different types of care (Lymer et al. 2009; Lymer et al. 2008); development of the SimObesity model to examine small area obesity among children (Procter, 2007); small area health-related conditions (Ballas et al. 2006a); the socio-economic impacts of major job gain or loss at the local level (Ballas et al. 2006b) and a range of other applications (Ballas et al. 2005a, 2005b; Clarke 1996).

Despite this progress, the spatial microsimulation field remains within its infancy and a range of important questions still need to be thoroughly reviewed and answered. In particular, one crucial area for further work is on-going validation of the reliability of the small area estimates produced from such models. As Voas and Williamson explain: 'The most challenging test of synthetic microdata … is how well they behave on topics that did not feature among the constraints' (2000, p. 360) [with 'constraints' being the small area Census benchmarks that the sample survey data have been reweighted to]. Voas and Williamson find that the simulated results from variables that have not been included within the Census benchmarks will generally only produce reasonable results if the actual distribution of those variables is close to the national norm or if the variables involved are highly correlated with the variables that were included within the Census benchmarks (2000, p. 360).

The recent NATSEM experience has reinforced this, with small area estimates of poverty appearing reasonable, presumably because poverty is highly correlated with benchmarks available within the Australian Census and used in the reweighting process (such as income and family type) (Chin et al. 2006a, 2006b). However, producing small area estimates of relatively rare disability states for the CAREMOD model proved much more difficult (Lymer et al. 2006). Similarly, 'most spatial microsimulation models are unsuitable for the prediction of variables that are affected considerably by external and localised factors, such as transport networks and public transport services, or the presence of a disproportionately large university or a single major employer in the region' (Ballas et al.

2005a, p. 14). Thus, for example, predicting the number of visits to a doctor within a small area by 'regionalising' a national health survey may greatly overstate the number of such visits for a particular small area that faces a chronic shortage of doctors. It is for such reasons that spatial microsimulation models in the future are likely to rapidly develop even more sophisticated approaches that utilise administrative data about actual supply or usage to adjust the first-round results from spatial microsimulation models[3].

Recently, Rahman (2009) reviews several methodologies for obtaining small area estimates, comparing statistical model-based approaches and microsimulation approaches He finds that spatial microsimulation analysis is robust and has particular advantages over other approaches, including the ability to further aggregate or disaggregate data into different spatial units; the capability to easily update and engage in further analysis; and through linking the model with a static microsimulation model, the ability to measure policy change at a small area level.

As this review has illustrated, spatial microsimulation is a fast-moving field, with growth being driven in large part by the intense interest of government departments and businesses in understanding much more about the characteristics of their clients within particular small areas. It is thus likely to be an area that continues to rapidly evolve.

This paper canvasses the steps required before a national sample survey can be reweighted to Census benchmarks.

2 PREPARATION OF SURVEY DATA

This section provides a brief overview of some of the key issues that need to be considered when choosing and preparing a sample survey for matching to Census small area benchmarks. While every country will have different national sample surveys and varying national population Censuses, the following is a checklist of issues that have emerged during NATSEM's Australian attempts to prepare various national sample surveys for reweighting to the Australian Census tables for small areas.

2.1 SCOPE OF THE SAMPLE

One of the most important things in reweighting the survey data to Census benchmarks is ensuring that the scope of the survey sample and the Census are the same or can be amended to be the same. While not universally the case, most of the national sample surveys conducted by the Australian Bureau of Statistics (ABS) only include the population in private dwellings within their scope. Thus, for example, the ABS Survey of Income and Housing (SIH) samples households and individuals resident in private dwellings and excludes those resident in non-private dwellings such as aged care and nursing homes, prisons, boarding schools and hospitals. If the sample survey file of households residing in

[3] This was done in NATSEM's HOUSEMOD model, where the initial housing estimates arising directly from the spatial microsimulation were adjusted using small area administrative data about Commonwealth Rent Assistance recipients (Phillips, 2006)

private dwellings is reweighted to Census small area targets that include those in non-private dwellings, then the results of the spatial microsimulation may be biased.

For example, suppose that we are trying to derive estimates of health status by age for each small area by reweighting a health sample survey file. If a particular small area contains five large nursing homes and we reweight to 'age by gender' Census totals for that small area, then we are likely to overstate the number of healthy over-70 year olds actually residing in that small area. This is because the relatively healthy over-70 year olds still living in their own homes in the sample survey will have been inflated to match the total number of over-70 year olds as shown in the Census results for that small area (a total which will include the relatively unhealthy over-70 year olds in nursing homes). Thus, it is important to ensure consistency between the population in scope in the relevant sample survey and the population in scope for the Census.

For some of our analysis, we do require information about people in non-private dwellings. In particular, we may be interested in people aged over 80, and a substantial proportion of these are in hospitals or nursing homes. To match survey data (which does not include people in non-private dwellings) to Census data (which does include people in non-private dwellings), NATSEM has developed an approach for imputing non-private dwelling records using the Census household sample file, which is a 1 per cent random sample from the Census containing unit record data. We then add these records onto the survey data, and benchmark to the Census small area data on people in non-private dwellings for small areas. More information on this process is in Section 2.7 below.

The other area where the scope of the Census and surveys do not match is people in non-classifiable households. These are households which are 'visitor only' households or where everyone is aged under 15, and which thus cannot be classified as standard household types on the Census. They are not included in any of the survey samples, so are not in the survey data; however they are on the Census data. To take these households into account, we specifically request Census benchmark tables that do not include non-classifiable households. More information on this process is in Section 3.3 below.

2.2 IMPUTATION OF CHILD RECORDS

Another problem that we have encountered with the survey data is that unit records are typically only created for persons aged 15 and over, and there are no individual records of children. This problem has been overcome somewhat by imputing the number of children in each household, through existing household information from the survey which tells us how many children in each age grouping reside in each household. Imputing child records onto the survey allows us to gain more accurate weights when benchmarking to Census data and it also allows child-focused research to take place, such as estimating the numbers of children in poverty at a small area level.

As detail regarding the number of children in each household is topcoded by the ABS for each age range (so the highest category given is 'Five or more children'), the variable for the total number of usual residents in each household, together with the total number of

children and adults, are used in order to re-estimate the number of children. Where it was determined that a household had an 'extra' child, this child was randomly assigned within the child age ranges available.

These dependent children were output as individual person level records, and assigned to their corresponding income units. Each child record was assigned a new individual identifier but the family identifier, income unit identifier, household identifier and the age group variables were retained. Relevant household and income unit variables were then merged onto the child-level dataset and appropriate person-level variables were assigned to each child. Child records were given the following values for each benchmark variable:

Age – as per age ranges in income unit file;

Sex – assigned randomly, based on the ratio between males and females in the Australian population according to the 2006 Census;

Study status – assumed studying full-time if 5 years and over;

Educational institution attending – assumed attending school if 5 years and over;

Level of highest non-school qualification – not applicable;

Marital status – not applicable;

Individual income – not applicable;

Labour force status – not applicable;

Occupation in main job – not applicable; and

Hours worked – not applicable.

2.3 CHOICE OF SURVEY DATA

The Australian Bureau of Statistics (ABS) Surveys of Income and Housing (SIH) were selected as the base surveys used to produce small area level household data for our latest spatial microsimulation models. (NATSEM's earlier efforts involved reweighting the 1998-99 Household Expenditure Survey to the 2001 Census data, and these projects are fully documented in Chin et al., 2006b and 2006c.) The SIH has rich detailed data on income and housing and is currently conducted every two years. In 2003-04 the SIH was conducted together with the Household Expenditure Survey (HES), and dwellings were selected through a stratified multistage cluster design. The sample excluded areas that were classified as "very remote" according to an index of remoteness produced by the ABS (ABS, 2007).

In order to maximise the sample size available for modelling we combined two SIH surveys. This also allowed the base data for the SpatialMSM model to be compatible with the base data for NATSEM's STINMOD static microsimulation model, which replicates the rules of the income tax, social security and family payments programs (Vu 2007, Lloyd

2007). This then has the significant advantage of providing the added potential of policy analysis at a small area level, if the research question calls for this (Harding et al. 2009b, Tanton et al. 2009a). For SpatialMSM versions pre-2009 we combined the 2003-04 SIH with the 2002-03 SIH while, for 2009 versions, we combined the 2003-04 and 2005-06 SIHs.

2.4 MATCHING VARIABLE DEFINITIONS IN SAMPLE SURVEY AND CENSUS

For the spatial microsimulation process to work correctly, the benchmarks used to reweight the sample survey data to Census small area targets must be defined in the same way in both datasets.

One key issue here involves matching variable definitions used in the sample survey and the Census small area tables. In some cases, this may require aggregating finer groups contained in either the Census or the sample survey to broader aggregations. For example, the sample survey may have eight categories of post-school qualifications while the Census may have only four. Careful reading of the documentation for both data sources is required to correctly aggregate the various categories so that they match exactly — a process which may, for example, ultimately end up with only two post-school qualification categories in both data sources.

A second issue is that variables that at first glance appear the same - for example, 'labour force status' - may be defined quite differently in the two data sources. For example, one data source may consider "Unemployed" as doing no hours of paid work per week; and another data source may define 'Unemployed' as those who receive unemployment benefits from the Government. In Australia, because both our data sources are from the same source (that is, the Australian Bureau of Statistics), variables are often defined the same way, as the ABS standard definitions applied across all surveys and the Census. However, this may be a more significant issue if the data were from different sources.

A third issue relates to categories such as 'not applicable' and 'not stated'. These are special categories used by the ABS for non-response (Not Stated) and out of scope families (so, for instance, households with a housing tenure of renters will have a 'not applicable' for amount of mortgage paid). It is also important to check the consistency of the population sub-groups covered within each variable.

2.5 UPRATING AND DEFLATING

Microsimulation modellers are accustomed to the concept of 'uprating', which typically involves adjusting monetary values collected within the sample survey to account for estimated price movements since the time of the survey, or anticipated future movements (Harding, 1996, p. 3). For example, a static microsimulation model that is trying to capture the 2008-09 tax and transfer systems may be built upon 2004 sample survey data, so that the earnings of employees shown in the survey data may need to be inflated by movements in average weekly earnings between 2004 and 2008-09.

In NATSEM's first SpatialMSM attempt, such uprating was required to match the 1998-99 Household Expenditure Survey values to those related monetary values in the 2001 Census (Chin et al. 2006c, p. 3). The 2001 Census, for example, might have a personal income category for gross income ranging between '$200 to $299' a week. To match to this, we first uprated various private income sources in the survey data by such factors as changes in average earnings, government cash benefits and the consumer price index to 2001 levels; then summed these to derive gross personal income. We then grouped these uprated survey incomes into bands that matched the Census income categories. Similar housing price inflators were used to create uprated data to match the 2001 Census categories for mortgages or rent paid.

NATSEM's second SpatialMSM model involved combining the 2002-03 and 2003-04 ABS Survey of Income and Housing Costs and then matching them to the 2001 Census (as the 2006 Census was not available at that time). This required us to deflate the 2002/03 and 2003/04 survey data to 2001 values. In essence, the monetary values captured in the sample surveys were adjusted by appropriate deflators to take them back to the levels prevailing at the time the Census was conducted. Then, once again, they were grouped into categories that matched the groupings in the Census. (It is worth noting that, after these matching variables were used in the reweighting process and the new small area weights had been derived for each household, these deflated values could then be dispensed with. Thus, for example, having used them to create the relevant small area household weights for 2001, we then used our standard STINMOD static microsimulation ageing techniques to 'age' the original 2002-03 and 2003-04 sample surveys up to a 2006 world, thus producing what were at the time highly current estimates of income poverty for 2006 - Tanton et al. 2009b; Lloyd, 2007).

Recent SpatialMSM efforts again involved uprating, this time combining the 2002-03 and 2003-04 income surveys for the SpatialMSM/08 versions, uprating monetary values to 2006, and then reweighting the surveys to the 2006 Census small area targets. As noted earlier, for the SpatialMSM/09 versions we combined the 2003-04 and 2005-06 SIHs and followed the same procedures outlined above.

Household, family and personal income have all been uprated (or deflated) using average weekly earning values, taken from ABS Cat No. 6302.0, full-time, adult, ordinary time earnings. Mortgage and rent values have been uprated using changes in prices, taken from ABS Cat No. 6401.0, consumer price index, all groups, original.

Because the surveys use a reference year, the inflating or deflating of both incomes and prices has been taken from the month that is closest to the mid-point of each survey (December). The Census is collected at a point in time (usually sometime in August), so when we inflate or deflate to match Census data, the inflator/deflator is taken to August.

For persons living in non-private dwellings (created from the 2001 Census sample file – see Section 2.1 above), the income values assigned had values that coincided with June 2007. These 2007 values were assigned based on uprating techniques from STINMOD. These values were deflated to values available for the closest month to August 2006.

2.6 MATCHING INCOME-SHARING UNITS

In many cases the sample survey and the Census may use the same income-sharing unit (the unit within which income is assumed to be shared). However, this is not always the case. In addition, it needs to be confirmed that a 'household' in the sample survey has the same meaning as a 'household' in the Census.

In the Australian case, NATSEM's STINMOD model uses a special 'social security' type income unit, which is a subset of the usual household income-sharing unit. In addition, some of the earlier ABS sample surveys used a nuclear family income unit definition, which was again a subset of the standard household income-sharing unit. In all such cases, the smaller income units had to be aggregated to household units before reweighting to the Census could occur.

2.7 IMPUTATION OF A NON-PRIVATE DWELLING POPULATION

Survey data does not often have information about persons living in non-private dwellings (which includes dwellings such as hospitals, boarding schools, prisons and nursing homes), whereas Census data does include this information. Given this inconsistency, information about non-private dwellings can either be deleted from or added to each data source in order to make them directly comparable.

Census data and persons in non-private dwellings

In pre-2008 versions of SpatialMSM, persons in non-private dwellings existed in the Census benchmark table - labour force by age by sex (a person-level benchmark table). However, recent special data requests from the ABS have made it possible to remove the NPD population from this important benchmark table, which means that we can have a direct population match to the survey data we are using. However, people living in non-private dwellings (NPDs) are often an important population to include in analyses. These people, whilst being a relatively small population compared to people in occupied private dwellings (OPD), are often recipients of income support, and are therefore of interest to policy makers. In 2006 there were around 320,000 persons usually resident in an NPD. This equates to approximately 3 per cent of Australia's population. Most of these people were residing in hotels, motels, boarding houses and homes for the aged and disabled. Others resided in hospitals, boarding schools and residence halls, prison and corrective institutions (see ABS 2003, table X45). Consequently, we have continued to include the NPD population in a separate benchmark table, as we wanted to be able to analyse this unique group of people.

To this effect, we have obtained a special table from the ABS which has the number of persons living in a non-private dwelling in each SLA for each of the twenty categories of non-private dwellings available. These categories have been amalgamated to the four categories available in the synthetic population that has been created to attach to the survey data for matching to the new Census benchmark table (described below). These categories are – 'Hotels, motels, boarding houses and private hotels'; 'Homes for the aged and disabled'; 'Hospitals (including psychiatric)'; and 'All other non-private dwellings

(including boarding schools and residence halls)'. This new benchmark table has also helped reconcile population differences that we were obtaining from our reweighting process, when people in NPDs were included in the person level age by sex by labour force benchmark table.

Creation of synthetic non-private dwelling data

As non-private dwelling records do not exist on the survey data, we needed to synthetically create such records. In NATSEM's first round of SpatialMSM, which involved reweighting the 1998-99 ABS Household Expenditure Survey unit record file to the 2001 Census tables, special records were created for individuals in non-private dwellings and such characteristics as country of birth, highest post-school qualification and marital status were imputed, based on detail available from the 1986 Census one per cent unit record file (Chin et al. 2006c, p. 24-27). These records were then attached to the survey unit record file. Currently we use a more recent NPD population, which is based on detail available from the 2001 Census one per cent sample file.

Both children and adults residing in NPDs were included in the sample, and only those persons classified as usual residents were included in the NPD population. Detail about these persons has been imputed from other known values – for example a single integer value of income has been imputed from the income range available in the sample file. Most other person-level detail was available, for example labour force status, age, number of hours worked, study status, type of educational institution attending and so on. Persons in NPDs received a value of zero for all household and family-level variables.

Each NPD record was given a weight of 100, given that the data has been derived from a 1 per cent random sample. This resulted in 199,500 adult (persons aged 15 and over) NPD records and 10,900 child (persons aged under 15 years) NPD records.

For more detail on the derivation of this population, please refer to NATSEM internal documentation – Abello (2005).

3 PREPARATION OF CENSUS DATA

3.1 DATA SOURCE AND SCOPE

Data from the 2006 ABS Census of Population and Housing were used as benchmarks - or constraints which the synthetic small estimates produced by the reweighting process must match. In Australia, the Census is conducted by the ABS once every five years, and information about the personal, family and dwelling characteristics of all Australians is collected. As mentioned above, the Census has the advantage (in contrast to sample surveys) of providing data at a high level of spatial disaggregation, however it provides less rich and detailed data than that available in surveys. The 2006 Census was conducted on the 8th August, 2006, and gives details of all people (including visitors) for each dwelling.

We have confined the scope of our benchmarks to only people who were in their usual place of residence on Census night. People classified as visitors, or in the Collection Districts of 'off-shore', 'shipping' and 'migratory' areas are excluded[4].

Personal characteristics in the Census data relate to the respondent's place of usual residence. In pre-2008 versions of SpatialMSM, which used the 2001 Census, the household and dwelling benchmark tables used the standard output from the ABS, which was data as enumerated on Census night; however, person-level variables used data for usual residence on Census night. In all ABS surveys, data is collected from the usual residents of a household, and visitors are excluded. In the 2006 Census, the standard output tables were changed so all tables were produced using usual residence data, and this is what we have used for all the benchmark tables in the later versions of SpatialMSM.

Most benchmark tables are multi-dimensional, as they are cross-tabulations of the variables that we want to benchmark to. Initially we planned to source the benchmark variables from the publicly available 2006 Census Basic Community Profile (BCP), and Expanded Community Profile (XCP). These BCP and XCP data provide characteristics of persons, families, households and dwellings at a geographically disaggregated level. However, for the most recent versions of SpatialMSM, most of the benchmark variables have been sourced via special tables requested from the ABS (see Table 1 for current SpatialMSM/09C benchmarks).

[4] Persons classified in off-shore, shipping and migratory Collection Districts are excluded during the reweighting process. Off-shore includes persons enumerated on an oil rig/drilling platform etc. Shipping includes persons enumerated on board vessels departing for an overseas port; and migratory covers all people who are in transit on long distance trains, buses and aircraft on Census Night (ABS 2006, p.171)

These tables have been sourced for the following reasons:

- The required benchmark table is not currently available either from either the BCP or XCP tables (an example is Benchmark Table 6, "Persons in non private dwellings");
- To align with definitions of variables in the SIH, that is, excluding households and persons who lived in "non-classifiable households" and excluding households and persons who lived in "being occupied rent-free" private dwellings; and
- To improve our results by requesting benchmark variables that are closely related to what we are trying to measure (for example, Benchmark Table 12, Gross equivalised weekly household income by age has substantially improved our small area estimates of older adults in poverty).

The list of benchmark tables used for our current spatial microsimulation model is shown in Table 1.

Table 1 Benchmark tables used for SpatialMSM/09C

No	Benchmark Table	Level	Sources
1	All household type	Household	XCP - X25
2	Age by sex by labour force status	Person	Special data request
3	Tenure by weekly household rent	Household	Special data request
4	Tenure by household type	Household	Special data request
5	Tenure by weekly household income	Household	Special data request
6	Persons in non-private dwellings	Person	Special data request
7	Monthly household mortgage by weekly	Household	Special data request
8	Dwelling structure by household family	Household	XCP - X25
9	Number of children aged under 15 usually	Household	Special data request
10	Number of adults usually resident in household	Household	Special data request
11	Weekly household rent by weekly household	Household	Special data request
12	Gross equivalised weekly household income by	Household	Special data request

3.2 BALANCING DATA

As the ABS randomises the data available at an SLA level to maintain a level of confidentiality, often benchmark table totals measuring the same population will not be the same. In NATSEM's previous spatial microsimulation models, a complicated and time-consuming process took place in order to align the Census benchmark total populations, as this was thought to improve the reweighting process and convergence. This process was called 'balancing' the tables. In 2007, some sensitivity analysis was conducted in order to determine if there was a significant bias in the results produced through using unbalanced data. A comparison of balanced and unbalanced results from the Australian Capital Territory (ACT) found that the practice of balancing data had little effect on the results - and, in some cases, the results obtained from the unbalanced data were closer to the true Census counts than those from the balanced data (NATSEM internal documentation – Tanton (2007). Currently, the Census benchmark tables used in SpatialMSM/09C are not balanced.

3.3 OTHER NON-CLASSIFIABLE HOUSEHOLDS

The Census data contains information about other non-classifiable households, which are not included in the survey data. 'Other non-classifiable households' are described as those households that contain no persons aged over 15 years; that the collector deemed occupied but was unable to make contact with any occupants; or where the information supplied on the Census form was inadequate. This discrepancy between the two data sources had to be corrected to make the data as consistent as possible. In earlier reweighting attempts, we would create a "pseudo" non-classifiable population by duplicating all household records on the SIH, thereby giving these households exactly the same characteristics as the classifiable households. As this was considered to be an inadequate solution, for the latest models we have obtained special request benchmark tables from the ABS that exclude non-classifiable households.

3.4 NOT STATED VALUES

Due to the nature of the collection of the Census data (non-interviewer assisted), the data contain fully and partially not-stated values. In the Survey of Income and Housing, any partial non-response (for example, on the Income question) is imputed, using available information from the household. Full non-response households are excluded from the sample, as no information is available from them.

In order to be able to benchmark the survey data to the Census tables, the not stated values on the Census tables were redistributed amongst other known categories. This redistribution was proportionate, based on the relative frequency of the true values of the known categories, so the not stated values were extrapolated out to other valid values.

3.5 NOT APPLICABLE VALUES

In the Census, not applicable values are values where the response to the question does not apply to the person or household, and so no response is required (ABS 2006). An example is mortgage loan repayments for renters. Not applicable categories can also include unoccupied private dwellings and migratory, offshore and shipping Collection Districts. We exclude relevant not applicable categories for each Census benchmark table.

4 DETAILED DESCRIPTION OF CHANGES MADE TO SURVEY AND CENSUS DATA TO MAKE THEM COMPARABLE

The choice of benchmarks to be used in the reweighting process is limited to the data available in both data sets - and will also be influenced by the socio-economic outcome variables required. For example, if we are trying to determine small area estimates of poverty, income is an important benchmark to include, as it is closely related to the outcome variable. Benchmarks are firstly created at the most detailed level available. However, during the reweighting process, the benchmark categories are often aggregated in order to improve the convergence as, the more categories within each benchmark, the more difficult the reweighting process becomes, and the higher the likelihood of non-converging areas.

This section provides a detailed description of the many measures that were taken in order to prepare the two ABS Surveys of Income and Housing Confidentialised Unit Record Files (CURFs) for reweighting to the 2006 Census SLA benchmark tables

4.1 PERSON LEVEL VARIABLES

This section discusses the recoding of the person-level variables that are common to both the SIHs and the Census.

Age

Three different age groupings have been derived to potentially be used as a benchmark variable, depending upon what level of detail is required and what will give the best reweighting results. However, due to age only being available in ten year categories from the Non-Private Dwelling file, if NPDs are to be included in the reweighting population, then ten year age groups must be used. It is important to note that age is topcoded in both the SIHs and the Census to maintain confidentiality. For the 2002-03, 2003-04 and 2005-06 SIHs, age is topcoded to 80 years. In the 2006 Census, age is topcoded at 115 years.

The first age variable is AGEP, and this is available in 19 groupings, ranging from two to five year groupings, with all persons aged 75 and above grouped together. The variable AGEP5 provides 5 year age groupings, with a final grouping of those aged 80+. The variable AGEP10 has six categories, ranging from 10 to 20 year age groupings, with the

final grouping of all those aged 65+, in order to match with the categories available in the NPD file.

Sex

A benchmark variable called SEXP has been derived from the SIH and Census. The sex of children has been imputed – see section 2.2 for more details.

Highest level of non-school qualification

Currently, there are two variables that can be used as benchmarks for the highest level of schooling completed – EDQUALP_06 and HQUALBC. EDQUALP_06 can only be used as a benchmark if the population with a post-school qualification is chosen, due to differences in the variable definitions provided in the 2006 Census, and the inability to establish a comparable definition between this variable and survey variables (the categories of "not applicable" and "still at school" cannot be matched). A new variable HQUALBC is also available, which provides a more detailed disaggregation of those with a non-school qualification. Again, this variable can only be used if each population contains only those persons with a non-school qualification.

Educational institution attending

Information about the educational institution being attended for all persons studying is available in both data sources and the benchmark variable EDUINT_06 has been constructed for all persons. The Census has greater detail available about the type of educational institution attending than the SIH does (e.g. Catholic, government, other non-government, full-time, part-time); and consequently these categories have been aggregated up in order to match those available in the SIH. There are five categories available for this benchmark variable, and those that are not studying receive a value of "0" which infers a 'not applicable' value. Children (those aged between 5 and 14) are assumed to be attending school.

Hours worked per week

The number of hours worked per week by a person aged 15 years and older is available on both the 2006 Census and relevant SIHs as a categorical variable. There are slight definitional differences between the hours worked variables available from each data source. The definition for number of hours worked in the SIHs is "Number of hours usually worked per week in main and second jobs", whereas in the Census this definition is "number of hours worked in the last week". Further, the working hours categories do not align exactly. More specifically, the SIH has the category 24-25 hours per week, whereas the Census categories are 16-24 and 25-34 hours per week, so a direct match cannot be achieved. Also, the Census has the hours range of 41-48 hours, whereas the SIH has the range 41-49 hours.

Children and persons not working receive a value of zero for this benchmark variable.

Due to the differences in definitions, we do not use this variable as a benchmark in our modelling.

Individual weekly income

Information on the total weekly gross personal income from all sources for a person 15 years and older is available as a continuous variable in the SIH. In the 2006 Census, total weekly gross income is only available in categories and, consequently, the SIH amounts are recoded to match the available Census categories. The benchmark variable is called ITINCP06.

As we are re-weighting the SIH information to individual income bands in the 2006 Census, the 2002-03 and 2003-04 incomes are uprated to the nearest month to August 2006 (the date of the Census) in order to make these amounts as comparable as possible. See section 2.5 for further detail about uprating factors.

Children receive a value of zero income value for this benchmark variable.

Occupation

Information about the occupation of a person aged 15 years and older is provided by the variable OCCBC in the SIH person file. Occupation is classified on the SIHs according to the Australian Standard Classification of Occupations (ASCO) 1997 Second Edition. However, in the 2006 Census, occupation is classified to the 2006 Australian New Zealand Standard Classification (ANZSCO).

These differences in classifications have been rectified somewhat by using a concordance provided by the ABS which allows the ANZSCO classifications to be recoded to ASCO (ABS 2008a).

We currently do not include occupation in our main benchmarks, but potentially could use it for helping to get small area estimates of wealth. Thus, it may be included in some of our models in the future.

Children receive a value of zero for this benchmark variable.

Labour force status

A combination of labour force status and hours worked are used in order to ascertain whether a person is working full- or part-time in each SIH. The variable labour force status (LFSCP) in each SIH also contains the categories unemployed and not in the labour force. The benchmark variable LFS has been created in order to match to the available Census categories. This information is available for those persons aged 15 and above only.

The 2006 Census has a new category for labour force status – 'employed away from work', which includes persons who stated they were employed but did not work any hours in the week prior to the Census. These persons have been re-distributed amongst the existing full-time and part-time employed persons on a pro-rata basis.

Children receive a value of zero for this benchmark variable.

4.2 HOUSEHOLD LEVEL VARIABLES

Number of dependents

The number of dependents under 15 years old in a household is available on the income unit files in the SIH within age ranges, and is topcoded. These values are accumulated to the household level and assigned to their corresponding household in order to match up with the Census household-level benchmark table.

There are three variables available that reflect the number of dependents: Number of dependents aged 0-4 years (DEPS0_4), Number of dependents aged 5-9 years (DEPS5_9), and Number of dependents aged 10-14 years (DEPS10_14).

The number of children aged under 15 (DEPS10_14) is currently used in the benchmarking, but the other variables could be used to get more reliable small area estimates for younger children.

Dwelling structure

Dwelling structure of a household is given the benchmark variable name of DWSTR_06. This variable has five categories, including a zero category for non-private dwellings. These categories include, 'separate house', 'semi-detached row, terrace or townhouse', 'flat, unit or apartment', and 'other dwelling' structure. It is possible to expand this variable with further detail about the number of stories the dwelling has. However, we decided that this level of detail was not required for the benchmarking.

Tenure type

The nature of housing occupancy (i.e. dwelling tenure type) is given the benchmark variable name of TENURE_06. The variable includes the following categories: 'Owner without a mortgage', 'Owner with a mortgage', 'Renting publicly', 'Renting privately' and 'Other tenure'.

As mentioned earlier, we requested a special table from the Census to separately identify households who live in 'rent free' accommodation. We were then able to move this category to the 'other tenure' tenure category to make it consistent with the SIH.

Household income

There are several weekly household income variables available which could be used in the benchmarking process. Each household income variable has been created from the original SIH household variable "INCTO1CH - Total current weekly HH income from all sources (pre 2003-04 SIH basis)". This variable has been uprated to match with the Census data as per the uprating technique described in section 2.5. The benchmark variables available are as follows:

a) Weekly household income (HHINCOME) – with twelve available categories, including a nil value for those in NPDs.

b) Gross weekly equivalised household income (HIED_06) – equivalised using the modified OECD equivalence scales. Equivalised income is derived by calculating the equivalence factor and then dividing gross income by the factor (ABS 2006, p.193). The equivalence factor is calculated by allocating points to each person in a household (1 point to the first adult, 0.5 points to each additional person who is 15 years and over, and 0.3 to each child under the age of 15) and then summing the points for all household members (ABS 2006, p.193). This variable has 9 available categories, and matches the categories available in the Census variable 'HIED'. Persons living in NPDs receive a value of zero for this variable.

c) Gross weekly household equivalised income quintiles (EquivincQnt)– this variable was constructed as per the gross weekly equivalised household income variable above. However, it is then put into quintiles based on equivalent numbers of households in each group. Persons living in NPDs receive a value of zero for this variable.

Household income is used in four of the current twelve benchmarks used to create SpatialMSM/09C. Three of these benchmark tables use gross weekly household income, and one benchmark table uses gross weekly equivalised household income.

Household composition and household family composition

The original SIH variable DCOMPH has been used to create two benchmark variables for the reweighting process – Household Composition and Household Family Composition.

a) Household composition (HHTYPE) has been created by collapsing the 33 available categories in the SIH into three broad household groups. These categories are 'Family household', 'Lone person household', and 'Group household'. This benchmark variable was originally used to match the Census table X12 (Tenure by household type). However, a special Census table for this benchmark was subsequently requested in order to remove those living "rent free". Persons living in NPDs receive a value of zero for this variable.

b) Household family composition (DCOMP_06) has been created by collapsing the 33 available categories in the SIH into six categories. These categories are as follows: 'Family household-couple family with children', 'Family household-couple family without children', 'Family household-one parent family', 'Family household-other family', 'Lone person household' and 'Group household'. This benchmark variable is used to match to Census table X25 (Dwelling Structure by Household Family Composition). Persons living in NPDs receive a value of zero for this variable.

Number of persons usually resident

Number of persons usually resident in the household (NPERSONS) has been constructed using the SIH variable PERSHBC. There are six available categories for this variable. This variable is topcoded to 6 or more persons in both the 2003-04 and 2005-06 SIHs. This benchmark variable was used in earlier versions of SpatialMSM. However, for later

versions, in order to gain more accurate counts of children, the variable has been split into children and adults. Persons living in NPDs receive a value of zero for this variable.

Number of children usually resident

Number of children usually resident in the household (NKIDS_06) has been derived from the SIH variable NUMU15BC, and has five available categories. The variable is top-coded to five or more children in both SIHs. A special request table was sought from the ABS in order to get Census small area data about the number of children usually resident in each household. Persons living in NPDs receive a value of zero for this variable.

Number of adults usually resident

Number of adults usually resident in the household (NADULTS_06) has been derived from the SIH variable NOMEMHBC, and has five available categories. The variable is top-coded to six or more adults in both SIHs. A special request table was sought from the ABS in order to get Census small area data about the number of adults usually resident in each household. Persons living in NPDs receive a value of zero for this variable.

Number of bedrooms in household

Number of bedrooms in a household (DNBED_06) has been derived from the original SIH variable NRBEDSCF and has five available categories. The variable is topcoded to five or more bedrooms. Persons living in NPDs receive a value of zero for this variable. This variable is currently not used for benchmarking, but could be used in future to achieve better small area estimates of household wealth.

Weekly rent payment

Total weekly rent payment of a household (RENT_06) is available in dollar amounts from the original SIH variable WKRENTCH. As we were re-weighting to weekly rents in the 2006 Census, these payments have been uprated using the uprating techniques described in Section 2.5. The uprated rent payments were available as a continuous variable, which was then recoded to ten categories to match the categories available in the Census. This benchmark variable was originally used to match the Census table X20-X23 (Gross household income by weekly rent) and B34 (Rent by landlord type). However, special Census tables for these benchmarks were subsequently requested in order to remove those living "rent free". Those not renting and persons living in NPDs receive a value of zero for this variable.

Monthly mortgage repayment

Information about weekly mortgage payments is available in dollar amounts from the original SIH using the variable TRPAY1CH (weekly mortgage repayments to purchase/build). A monthly mortgage repayment was calculated by multiplying this variable by four. As we were re-weighting the SIH information to the monthly mortgages in the 2006 Census, these payments were uprated using the uprating technique described in Section 2.5. The uprated mortgage payments were available as a continuous variable,

which was then recoded to eight categories to match the mortgage dollar ranges in the 2006 Census XCP table.

4.3 FAMILY LEVEL VARIABLES

Family type

Information about the family type of a person is provided by the variable FAMTYPE_06 and is derived from the variable FAMTYPE in the original SIH person file, and has 16 categories. These categories were reduced to five, including a zero not applicable category for those not residing in a family, and those living in NPDs. The variable is only applicable for those who are considered to be residing in a family, given the current ABS definition. This definition is as follows:

"Two or more people, one of whom is at least 15 years of age, who are related by blood, marriage (registered or de facto), adoption, step or fostering, and who usually live in the same household. A separate family is formed for each married couple, or for each set of parent–child relationships where only one parent is present." (ABS 2008b, p. 56)

Family income

Weekly family income (FAMINC) is created by summing all of the individual family members' total gross weekly incomes from all sources (INCTOTCP). This variable (INCTOTCP) is uprated using the uprating technique described in Section 2.5. This variable was then re-coded in line with the family weekly income classification in the 2006 Census tables. The variable has 16 categories, including a value of zero for those living in NPDs and those persons that do not reside within a family, as per the definition given above.

4.4 LIST OF POSSIBLE BENCHMARKS

A full list of benchmark variables, with the categories available, is shown in Table 2. Note that not all these variables are used in every version of SpatialMSM; the choice of variables benchmarked to depends on the outcome variables that we are trying to estimate (such as poverty or housing stress).

Table 2 Benchmark variables and categories available for reweighting process

Benchmark variable name created from 2003-04 and 2005-06 Surveys of Income and Housing to link with 2006 Census.	Categories available	
AGEP	Age of the person	11=35-39 years
	1=0-2 years	12=40-44 years
	2=3-4 years	13=45-49 years
	3=5-9 years	14=50-54 years
	4=10-14 years	15=55-59 years
	5=15-17years	16=60-64 years
	6=18-19 years	17=65-69 years
	7=20-22 years	18=70-74 years
	8=23-24 years	19=75+
	9=25-29 years	
	10=30-34 years	
AGEP5 - Age of the person (5 year groups)	1=0-4 years	10=45-49 years
	2=5-9 years	11=50-54 years
	3=10-14 years	12=55-59 years
	4=15-19 years	13=60-64 years
	5=20-24 years	14=65-69 years
	6=25-29 years	15=70-74 years
	7=30-34 years	16=75-79 years
	8=35-39 years	17=80+
	9=40-44 years	
AGEP10 - Age of the person (concatenated age groups)	1=0-14 years	4 = 35-54 years
	2= 15-24 years	5 = 55-64 years
	3 = 25-34 years	6 = 65+ years
EDQUALP_06 - Level of highest non-school qualification	0=not applicable	2=Higher/bachelor degree, postgraduate diploma
	1=still at school	3=Other post school qualifications
HQUALBC Level of highest non-school qualification	1 = Postgraduate degree, graduate diploma/graduate certificate	4 = Certificate III / IV
	2 = Bachelor degree	5 = Certificate I / II
	3 = Advanced diploma / diploma	6 = Certificate not further defined
		7 = No non-school qualification
		8 = Level not determined
EDUINT_06 - Educational institution attending	0 = not applicable	2 = TAFE
	1 = School	3 = University
		4 = other

Benchmark variable name created from 2003-04 and 2005-06 Surveys of Income and Housing to link with 2006 Census.	Categories available	
HRSJOB_06 - Number of hours usually worked per week in main and second jobs	0=zero hours 1=1-15 hours 2=16-23 hours 3=24-34 hours	4=35-39 hours 5=40 hours 6=41 to 49 hours 7=50+ hours
ITINCP06 - Individual weekly income	0=not applicable (under 15 years old) 1=nil/negative income 2=$1-$149 3=$150-$249 4=$250-$399	5=$400-$599 6=$600-$799 7=$800-$999 8=$1000-$1299 9=$1300-$1599 10=$1600-$1999 11=$2000+
LFS - Labour Force Status of Person	0=Not applicable (children) 1=Employed - full-time 2=Employed - part-time	3=Unemployed 4=not in labour force 99=Unemployed & not in the labour force, aged 65+
OCCP_06 - Occupation in main job	0=Not applicable 1=Managers and Administrators 2=Professionals 3=Associate Professionals 4=Tradespersons and Related Workers 5=Advanced Clerical and Service Workers	6=Intermediate Clerical, Sales and Service Workers 7=Intermediate Production and Transport Workers 8=Elementary Clerical, Sales and Service Workers 9=Labourers and Related Workers 10=Inadequately described
SEXP - Sex of the person	1=male	2=female
STUDY - Education status of the person	0=not applicable 1=full-time student	2=part-time student 3=not studying
DEPS0_4	Number of dependents aged 0-4 years	
DEPS5_9	Number of dependents aged 5-9 years	

Benchmark variable name created from 2003-04 and 2005-06 Surveys of Income and Housing to link with 2006 Census.	Categories available	
DEPS10_14	Number of dependents aged 10-14 years	
DNBED_06 - Number of bedrooms	0=NPD 1=No bedrooms/1 bedroom 2=two bedrooms	3=three bedrooms 4=four bedrooms 5=five or more bedrooms
DWSTR_06 - Dwelling structure	0=NPD 1=separate house 2=semi-detached	3=flat, unit, apartment 4=other dwelling structure
TENURE_06 - Tenure type	0=NPD 1 = owner without a mortgage 2=owner with a mortgage	3=rent - Public 4=rent – private 5 = other
HHINCOME - Weekly household income	0=NPD 1=Nil or negative income 2=$1-$149 3=$150-$249 4=$250-$499 5=$500-$649	6=$650-$799 7=$800-$999 8=$1,000-$1,399 9=$1,400-$1,699 10=$1,700-$1,999 11=$2,000+
HIED_06 - Gross household equivalised income	0= NPD 1=Negative/nil income 2=$1-$149 3=$150-$399 4=$400-$799	5=$800-$1299 6=$1300-1599 7=$1600-$1999 8=2000+
EquivincQnt - Gross household equivalised income quintiles	0=NPD 1='Bottom Quintile' 2='Second Quintile'	3='Third Quintile' 4='Fourth Quintile' 5='Highest Quintile'
DCOMP_06 - Household family composition	0=NPD 1=family household-couple family with children 2=family household- couple family without children	3=family household-one parent family 4=family household-other family 5=lone person household 6=group household
HHTYPE - Household composition	0=NPD 1=family household 2=lone person household 3=group household	

Benchmark variable name created from 2003-04 and 2005-06 Surveys of Income and Housing to link with 2006 Census.	Categories available	
NPERSONS - Number of usual residents in household	0=NPD 1=one residents 2=two residents 3=three residents	4=four residents 5=five residents 6=six residents
NADULTS_06 – Number of adults usually resident in household	0=NPD 1 = 1 adult 2 = 2 adults	3 = 3 adults 4 = 4 adults 5 = 5 or more adults
NKIDS_06 – Number of children usually resident in household	99= NPD 0= 0 kids 1 = 1 kid	2 = 2 kids 3 = 3 kids 4 = 4 or more kids
RENT_06 – Weekly Rent payments for household	0=not applicable (not renting) + NPD 1=$0-$49 2=$50-$99 3=$100-$139 4=$140-$179	5=$180-$224 6=$225-$274 7=$275-$349 8=$350-$449 9=$450-$549 10=$550+
MORT_06 - Weekly mortgage repayments to purchase/build	0=not applicable + NPD 1=$1–$149 2=$150–$399 3=$400–$649 4=$650–$949	5=$950–$1399 6=$1400–$1999 7=$2000–$2999 8=$3000+
NPDTYPE - Type of non-private dwelling	0=private dwelling 1 = Hotels, motels, boarding houses and private hotels 2 = Homes for the aged and disabled	3 = Hospitals (including psychiatric) 4 = All other non-private dwellings (including boarding schools and residence halls)
FAMTYPE_06 - Family type	0=not applicable 1=couple family with children 2=couple family without children	3=one parent family 4=other family
FAMINC - Weekly family income	1=Nil or negative income 2=$1-$149 3=$150-$249 4=$250-$349 5=$350-$499 6=$500-$649 7=$650-$799	9=$1,000-$1,199 10=$1,200-$1,399 11=$1,400-$1,699 12=$1,700-$1,999 13=$2,000-$2,499 14=$2,500-$2,999 15=$3,000+

Benchmark variable name created from 2003-04 and 2005-06 Surveys of Income and Housing to link with 2006 Census.	Categories available
	8=$800-$999

5 CONCLUSION

NATSEM is engaged in a long-term project to create synthetic small area household microdata that are sufficiently reliable to be used by policy makers in such applications as needs-based planning, the analysis of disadvantage and modelling the spatial impacts of policy change. The latest version of NATSEM's spatial microsimulation model, SpatialMSM/09E, involved combining the 2003-04 and 2005-06 income surveys, uprating them to 2006, and then reweighting them to the 2006 Census small area targets. Currently (2010), work is underway to combine the latest survey data – the 2007/08 SIH with the 2006 Census data, for newer versions of SpatialMSM.

This paper details the specific measures taken to prepare the sample survey data for subsequent reweighting to the Census benchmark small area tables and the creation of NATSEM's spatial microsimulation model SpatialMSM. Numerous papers and online maps have been produced from SpatialMSM, and much of this output can be found on NATSEM's website.

REFERENCES

Abello, A., and Percival, R., 2005, 'Including the non-private dwelling population in STINMOD', NATSEM internal paper.

ABS 2006, 'Census Dictionary Australia 2006 (Reissue)', ABS Cat No. 2901.0, Canberra.

There's a reference to ABS(2007) in the text as well

ABS 2008a, Census of Population and Housing: Link Between Australian Standard Classification of Occupations (ASCO) Second Edition and Australian and New Zealand Standard Classification of Occupations (ANZSCO), ABS Cat. No. 1232.0, Canberra.

ABS 2008b, Survey of Income and Housing, Confidentialised Unit Record Files, Technical Manual, Australia, 2005-06 Second Edition, ABS Cat No. 6541.0, Canberra

Ballas, D., Rossiter, D., Thomas, B., Clarke, G.P. and Dorling, D. 2005a, Geography Matters: Simulating the Local Impacts of National Social Policies, York, Joseph Rowntree Foundation.

Ballas, D., Clarke, G.P. and Weimers, E, 2005b. 'Building a Dynamic Spatial Microsimulation Model for Ireland' , Population, Space and Place, vol. 11, pp 157-172.

Ballas, D., Clarke, G., Dorling, D., Rigby, J., and Wheeler, B., 2006a, "Using geographical information systems and spatial microsimulation for the analysis of health inequalities" Health Informatics Journal, 12 65-79.

Ballas, D., Clarke, G.P. and Dewhurst, J., 2006b, 'Modelling the Socio-Economic Impacts of Major Job Loss or Gain at the Local level: A Spatial Microsimulation Framework, Spatial Economic Aalysis, vol. 1, no. 1, June, pp 127-146,

Brown, L., Abello, A., Phillips, B. and Harding, A., 2004, 'Moving Towards an Improved Micro-Simulation Model of the Australian Pharmaceutical Benefits Scheme', Australian Economic Review, vol. 37, no. 1, pp. 41-61.

Chin, S.F., Harding, A., Lloyd, R., McNamara, J., Phillips, B. and Vu, Q.N., 2005, 'Spatial microsimulation using synthetic small-area estimates of income, tax and social security benefits', Australasian Journal of Regional Studies, vol. 11, no. 3, pp. 303-336. (See also NATSEM Online Conference Paper CP 0516, which contains maps).

Chin, S.F., Harding, A. and Tanton, R., 2006a,'A Spatial Portrait of Disadvantage: Income Poverty by Statistical Local Area in 2001', paper presented at the ANZRSAI Conference on Heritage and Regional Development, Beechworth, Victoria, 26-29 September.

Chin, S.F. and Harding, A., 2006b. 'Regional Dimensions: Creating Synthetic Small-area Microdata and Spatial Microsimulation Models'. Technical Paper no. 33, NATSEM, University of Canberra , April.

Chin, S.F., Harding, A. and Bill, A., 2006c, Regional Dimensions: Preparation of the 1998-99 Household Expenditure Survey for Reweighting to Small-area Benchmarks, Technical Paper no. 34, NATSEM, University of Canberra.

Clarke, G.P., (ed) 1996, Microsimulation for Urban and Regional Policy Analysis,

Elazar, D. and Conn, L. 2005, Small area estimates of Disability in Australia, Canberra: ABS, Publication no. 1351.0.55.006.

EURAREA Consortium, August 2004, Enhancing Small Area Estimation Techniques to meet European Needs, cited on www.statistics.gov.uk/eurarea/default.asp

Gupta, A. and Kapur, V. (eds), 2000, Microsimulation in Government Policy and Forecasting, North Holland, Amsterdam.

Gupta, A. and Harding, A., (eds), 2007, Modelling Our Future: Population Ageing, Health and Aged Care,, International Symposia in Economic Theory and Econometrics, North Holland, Amsterdam.

Harding, A., (ed), Microsimulation and Public Policy, Contributions to Economic Analysis Series, Amsterdam, North Holland, 1996.

Harding, A., Abello, A., Brown, L. and Phillips, B., (2004), 'Distributional Impact of Government Outlays on the Australian Pharmaceutical Benefits Scheme in 2001-02', Economic Record, vol. 80, no. S1, pp. S83-S96.

Harding, A and Gupta, A, 2007a, 'Introduction and Overview', in Harding, A and Gupta, A (eds), Modelling Our Future: Population Ageing, Social Security and Taxation (eds), International Symposia in Economic Theory and Econometrics, North Holland, Amsterdam.

Harding, A and Gupta, A, 2007b, Modelling Our Future: Population Ageing, Social Security and Taxation (eds), International Symposia in Economic Theory and Econometrics, North Holland, Amsterdam.

Harding, A., Vu, Q.N., Payne, A. and Percival, R., (2009a), 'Trends in Effective Marginal Tax Rates in Australia from 1996-97 to 2006-07', The Economic Record, vol. 85, no. 271, pp. 449-461.

Harding, A., Vu, Q.N., Tanton, R. and Vidyattama, Y., (2009b), 'Improving Work Incentives and Incomes for Parents: The National and Geographic Impact of Liberalising the Family Tax Benefit Income Test', The Economic Record, vol. 85, Special Issue, pp. 48–58.

Immervoll, H., Levy H., Lietz, C., Mantovani, D., O'Donoghue, C., Sutherland, H., and Verbist, G. 2006, (eds), 'Household incomes and redistribution in the European Union; quantifying the equalizing properties of taxes and benefits' in Papadimitriou, D. B. The Distributional Effects of Government Spending and Taxation, Palgrave, MacMillan, pp. 135-165.

Lloyd, R, 2007, 'STINMOD: Use of a Static Microsimulation Model in the Policy Process in Australia', in Harding, A. and Gupta, A., 2007, Modelling Our Future: Population Ageing, Social Security and Taxation (eds), International Symposia in Economic Theory and Econometrics, North Holland, Amsterdam.

Lymer, S., Brown, L., Harding, A., Yap, M., Chin, S.F. and Leicester, S., 2006, Development of CareMod/05, Technical paper no. 32, NATSEM, University of Canberra.

Lymer, S., Brown, L., Yap, M. and Harding, A. (2008), 'Regional disability estimates for New South Wales in 2001 using spatial microsimulation'. Applied Spatial Analysis and Policy, vol. 1, no. 2, pp 99-116.

Lymer, S., Brown, L., Harding, A. and Yap, M., 2009, 'Predicting the need for aged care services at the small area level: the CAREMOD spatial microsimulation model'. International Journal of Microsimulation. vol. 2, no. 2, Autumn, pp. 27-42.

McNamara, J., Tanton, R., and Phillips, B., 2007, 'The regional impact of housing costs and assistance on financial disadvantage', report for the Australian Housing and Urban Research Institute, RMIT-NATSEM Research Centre, AHURI Final Report No. 109.

Mitton, L., Sutherland, H. and Weeks, M., (eds) 2000, Microsimulation Modelling for Policy Analysis, Cambridge University Press, Cambridge.

Orcutt, G., Merz, J. and Quinke, H, 1986, Microanalytic Simulation Models to Support Social and Financial Policy, North Holland, New York.

Pfeffermann D. 2002. Small Area Estimation- New Developments and Directions. International Statistical Review, vol. 70, pp. 125 – 143.

Phillips, B., Kelly, S., 2006, 'Housemod: A regional Microsimulation Projections Model of Housing in Australia', paper presented at the Australian Housing Research conference, 19-21 June.

Procter, K., 2007, 'How where we live influences obesity: a geo-demographic classification of obesogenic environments using spatial microsimulation modelling', paper presented at the American Association of Geographers, San Francisco, USA, 17-21 April.

Rahman, A., 2009, "Small Area Estimation Through Spatial Microsimulation Models: Some Methodological Issues', Paper presented at the 2nd International Microsimulation Association Conference, Ottawa, Canada, 8-10 June.

Rao, J.N.K., 2003, Small Area Estimation, New Jersey, John Wiley & Sons, Inc.

Tanton, R. 2007, 'To balance or not to balance?', NATSEM internal paper.

Tanton, R., Vidyattama, Y., McNamara, J., Vu, Q.N., and Harding, A., 2009a 'Old, Single and Poor: Using Microsimulation and Microdata to Analyse Poverty and the Impact of Policy Change among Older Australians', Economic Papers: A journal of applied economics and policy, vol.28, no. 2, pp. 102– 120.

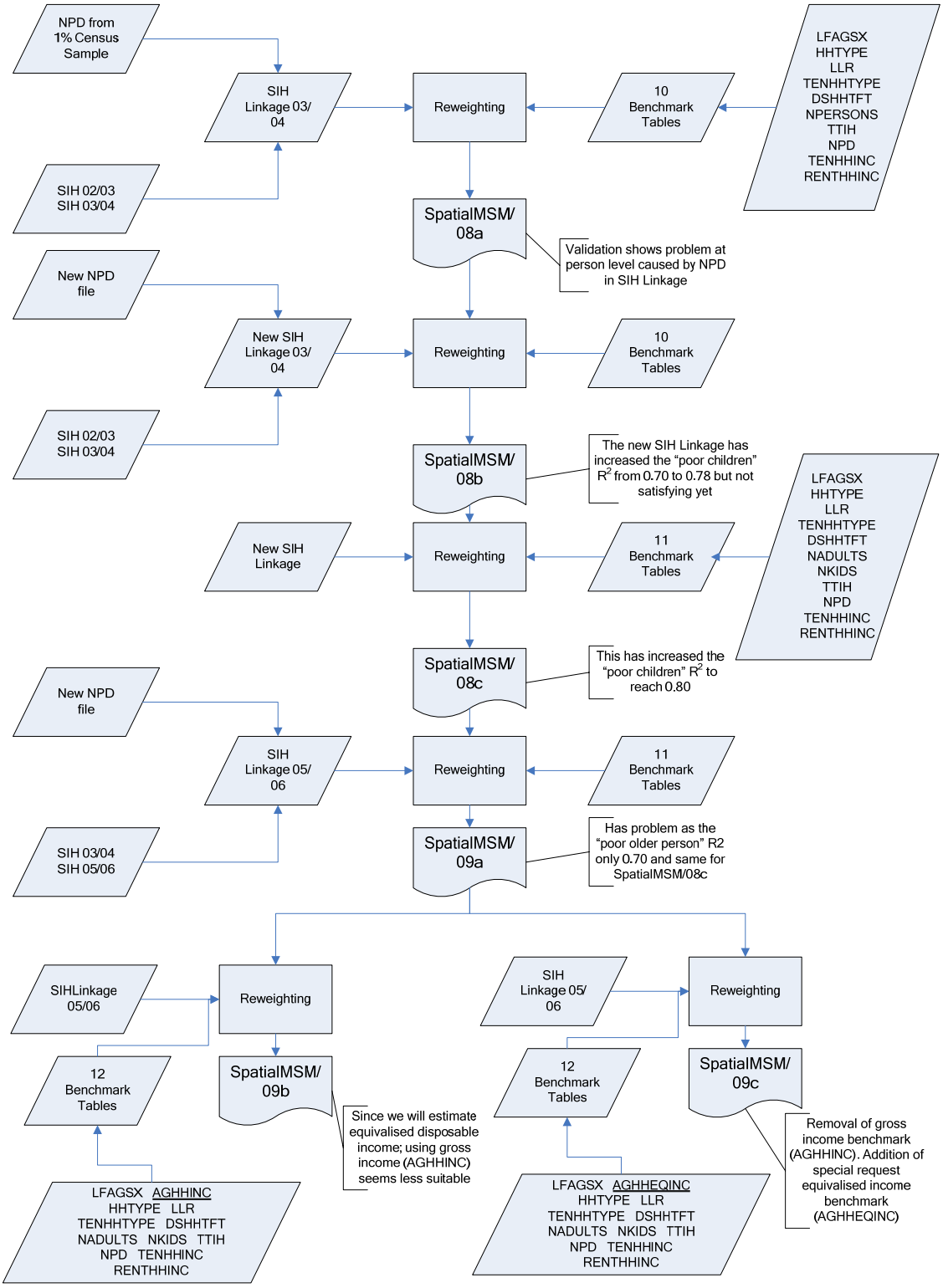
Tanton, R., McNamara, J., Harding, A. and Morrison, T., 2009b, 'Rich suburbs, poor suburbs? Small area poverty estimates for Australia's eastern seaboard in 2006'. In Zaidi, A., Harding, A., and Williamson, P., (eds) New Frontiers in Microsimulation Modelling, Ashgate, London (earlier version on NATSEM website as Conference Paper No 108).

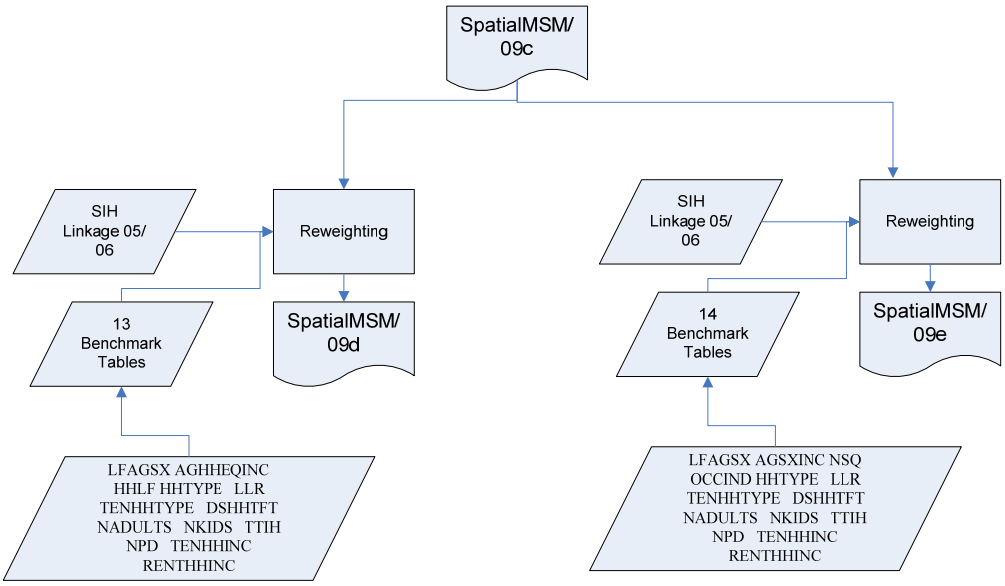
Voas, D. and Williamson, P., 2000, 'An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata', International Journal of Population Geography, vol. 6, pp 349-366.

Vu, Q.N. 2007, 'STINMOD Interface User Guide V06b', National Centre for Social and Economic Modelling, University of Canberra. Available from: http://www.canberra.edu.au/centres/natsem/publications?sq_content_src=%2BdXJsPWh0dHAlM0ElMkYlMkZ6aWJvLndpbi5jYW5iZXJyYS5lZHUuYXUlMkZuYXRzZW0lMkZpbmRleC5waHAlM0Ztb2RlJTNEcHVibGljYXRpb24lMjZwdWJsaWNhdGlvbiUzRDExMTMmYWxsPTE%3D Accessed 16 March 2010.

Williamson, P., Birkin, B., and Rees, PH. 1998, 'The estimation of population microdata by using data from small area statistics and samples of anonymised records', Environment and Planning A, pp. 785-816.

APPENDIX A


NPD from 1% Census Sample
SIH Linkage 03/ 04
Reweighting
10 Benchmark Tables
LFAGSX HHTYPE LLR TENHHTYPE DSHHTFT NPERSONS TTIH NPD TENHHINC RENTHHINC
SIH 02/03 SIH 03/04
SpatialMSM/ 08a
Validation shows problem at person level caused by NPD in SIH Linkage
New NPD file
New SIH Linkage 03/ 04
Reweighting
10 Benchmark Tables
SIH 02/03 SIH 03/04
SpatialMSM/ 08b
The new SIH Linkage has increased the "poor children" R^2 from 0.70 to 0.78 but not satisfying yet
New SIH Linkage
Reweighting
11 Benchmark Tables
LFAGSX HHTYPE LLR TENHHTYPE DSHHTFT NADULTS NKIDS TTIH NPD TENHHINC RENTHHINC
SpatialMSM/ 08c
This has increased the "poor children" R^2 to reach 0.80
New NPD file
SIH Linkage 05/ 06
Reweighting
11 Benchmark Tables
SIH 03/04 SIH 05/06
SpatialMSM/ 09a
Has problem as the "poor older person" R2 only 0.70 and same for SpatialMSM/08c
SIHLinkage 05/06
Reweighting
SIH Linkage 05/ 06
Reweighting
12 Benchmark Tables
SpatialMSM/ 09b
Since we will estimate equivalised disposable income; using gross income (AGHHINC) seems less suitable
12 Benchmark Tables
SpatialMSM/ 09c
Removal of gross income benchmark (AGHHINC). Addition of special request equivalised income benchmark (AGHHEQINC)
LFAGSX AGHHINC HHTYPE LLR TENHHTYPE DSHHTFT NADULTS NKIDS TTIH NPD TENHHINC RENTHHINC
LFAGSX AGHHEQINC HHTYPE LLR TENHHTYPE DSHHTFT NADULTS NKIDS TTIH NPD TENHHINC RENTHHINC



Benchmark Table Details

Number	Benchmark	Name
1	Age by sex by labour force status	LFAGSX
2	Age by weekly household equivalised income	AGHHEQINC
3	Age by weekly household gross income	AGHHINC
4	Age by sex by weekly person income	AGSXINC
5	Non School Qualification	NSQ
6	Occupation by Industry	OCCIND
7	Total number of households by dwelling type (Occupied private dwelling/Non private dwelling)	HHTYPE
8	Tenure by weekly household rent	LLR
9	Tenure by household type	TENHHTYPE
10	Dwelling structure by household family composition	DSHHTFT
11	Number of persons usually resident in household	NPERSONS
12	Number of adults usually resident in household	NADULTS
13	Number of children usually resident in household	NKIDS
14	Monthly household mortgage by weekly household income	TTIH
15	Persons in non-private dwelling	NPD
16	Tenure type by weekly household income	TENHHINC
17	Weekly household rent by weekly household income	RENTHHIC