NATSEM
at the University of Canberra

Small Area Social Indicators for the Indigenous Population: Synthetic data methodology for creating small area estimates of Indigenous disadvantage

NATSEM Working Paper 13/24

Yogi Vidyattama
Robert Tanton
Nicholas Biddle

September 2012

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.

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

Phone + 61 2 6201 2780
Fax + 61 2 6201 2751
Email natsem@natsem.canberra.edu.au
Website www.natsem.canberra.edu.au

Contents

Author note

Robert Tanton is Research Director of the Regional and Urban Modelling team at NATSEM at the University of Canberra.

Yogi Vidyattama is a Senior Research Fellow at NATSEM and part of the Regional and Urban Modelling.

Nicholas Biddle is an applied behavioural scientist and a Fellow at the Centre for Aboriginal Economic Policy Research (CAEPR) at the Australian National University (ANU).

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.

The citation for this paper is: Vidyattama,Y., Tanton,R., and Biddle, N. (2013), 'Small Area Social Indicators for the Indigenous Population: Synthetic data methodology for creating small area estimates of Indigenous disadvantage', NATSEM Working Paper 2013/24, NATSEM: Canberra

Abstract

The lack of data on how the social condition of Indigenous people varies throughout Australia has created difficulties in allocating government and community programs across Indigenous communities. In the past, spatial microsimulation has been used to derive small area estimates to overcome such difficulties. However, for previous applications, a record unit file from a survey dataset has always been available on which to conduct the spatial microsimulation. For the case of indigenous disadvantage, this record unit file was not available due to the scarcity of the Indigenous population in Australia, and concerns from the ABS about confidentialising the file. This study offers a solution to this problem by proposing the building of a synthetic unit record file with observations that sum to the population totals from the actual survey file. A spatial microsimulation approach is then applied to this synthetic unit record file and the results are validated.

1 Introduction

A key input into the development of public policy in Australia is the geographic distribution of socioeconomic outcomes. Evidence-based assessment of need has to take into account not only where people live, but also their characteristics. This has led to the widespread use of the Socio-economic Indexes for Areas (SEIFA), developed by the ABS and updated each five-yearly census. However, these indexes do not capture the differences in Indigenous disadvantage across the country. Evidence has shown that an index of Indigenous disadvantage gives different results to a general index of disadvantage like SEIFA, primarily due to the fact that non-Indigenous advantage masks a great deal of Indigenous disadvantage (see Kennedy and Firman, 2004). Furthermore, there are different components of wellbeing that need to be considered for Indigenous people that are not captured by the Census.

The need for different social indicators at a local level has caused small area estimation techniques to flourish (Ghosh and Rao, 1994; Pfeffermann, 2002). These techniques include methodologies that are based on the regression of predictors (Elbers et al 2003, Pratesi and Salvati 2008, Fabrizi et al 2012) and the reweighting of survey data known as spatial microsimulation modelling (Caldwell et al., 1998; Ballas et al., 2005 Nakaya et al., 2007; Vidyattama et al., 2011). In Australia, the spatial microsimulation method has increasingly been used in the past to derive small area estimates of a range of economic and social indicators, as well as estimating the impact of government policy and the need for government services at a small area level (Harding and Tanton, 2011). Examples of this work include simulating the small area impact of changes in income taxes and cash transfers (Chin et al., 2005; Harding et al., 2009; Tanton et al., 2009); the need for different types of aged care (Lymer et al., 2008); the retirement saving by gender of those who have just retired (Vidyattama et al., 2011) as well as measuring the distribution of trust (Hermes and Poulsen, 2012).

All of these small area estimation techniques bring together survey data that contains a specific variable but does not have enough observations to represent a small area with census or administrative data with enough observations in a small area to derive a reliable estimate. This is especially true for spatial microsimulation methods that require a unit record survey file.

Unfortunately, in some cases, this unit record survey file is not available, maybe due to confidentiality reasons; or because the owner of the dataset cannot release it. This is the case with the National Aboriginal and Torres Strait Islander Social Survey (NATSISS) from the Australian Bureau of Statistics (ABS). This file contains information that is considered sensitive by the ABS and is therefore not released to the public.

This work reports on the development of a method that allows spatial microsimulation to be conducted when survey unit record data is not available. The model used for this paper is one variant of a spatial microsimulation model that uses a generalised regression reweighting method to reweight survey data to known small area totals from the Census.

Following the introduction, the structure of this paper is as follows. Section two presents the data and methodology that will be used to construct the data base. Section three goes through the process of building the synthetic unit record data; Section four describes the application of the spatial microsimulation process; and Section five concludes the paper.

2 Data and Methodology

2.1 Data Sources

As mentioned in the introduction, the estimation process requires two main databases. The first is the census of population and housing. The Australian Bureau of Statistics (ABS) conducts a nationwide Census to obtain a count of the number of people in Australia, their individual characteristics and their dwelling characteristics every five years. The latest two censuses were on 8 August 2006 and 9 August 2011, respectively. Both capture more than 20 million people. This study used those two census datasets.

Indigenous people in the census are identified by the question that asks whether a person is of Aboriginal and/or Torres Strait Islander decent. Among the 20 million people in Australia, around 455,000 in 2006 and 548,000 in 2011 identified themselves as Aboriginal or Torres Strait Islander. Between the 2006 and 2011 census, the increase in the number of Indigenous people was 4.10 per cent annually, higher than the total population. The high growth rate in the number of Indigenous people does not come only from population growth, but may also be due to an improvement in Indigenous enumeration and self-identification following the ABS Indigenous Enumeration Strategy (IES). An evaluation conducted of the IES after the 2006 census has led to continuous and extended Indigenous community engagement and involvement in the 2011 census (ABS 2011a, Morphy et al., 2007).

The census data in Australia contain characteristics such as age and sex, cultural and language diversity, disability and carer status, children and childcare, employment and income, education and qualifications as well as relationship. Some of the questions in the census also ask about the condition of the family and the household such as the number of children, family composition, dwelling characteristic and tenure and household income. Most of the questions are multiple choice with few open-ended questions. Some of the numeric data such as income are collected based on income ranges. The main advantage of using census data for this study is the geographical information it contains.

Two different Census's were used for this work. The synthetic Indigenous dataset used the 2006 Census data, while the spatial microsimulation process used the 2011 Census data. The reason for this was that the 2011 Census data was not available when the synthetic Indigenous dataset was created.

The ABS currently uses the Australian Statistical Geography Standard (ASGS) as the main structure of census data dissemination. This has replaced the use of the ASGC since 2011.In the ASGS, Statistical Areas Level 1 (SA1s) are designed to have an average population of about 400 while Statistical Areas Level 2 (SA2s) have an average population of about 10,000, with a minimum population of 3,000 and a maximum of 25,000.

Besides the ASGS classification, there are additional area classifications in the census based on other known structures. The Indigenous Structure is one of these. This area is based on the existence of known Indigenous communities that should be represented by the area.

The second database used for this work is the National Aboriginal and Torres Strait Islander Social Survey (NATSISS) 2008. The main aim of this survey is to capture health, education, culture and labour force participation of Indigenous people. In addition, the 2008 survey also captured population characteristics such as age and sex, social capital such as social contact and networks, life experiences such as bullying and discrimination, housing and mobility, transport, information technology (i.e. computer, internet, telephone) as well as safety and experiences of crime in the home.

The ABS conducted the 2008 survey using face to face interviews between August 2008 and April 2009. There were 13,307 observations in the 2008 survey to represent around 518,000 Indigenous people across Australia. Although most of the Indigenous population lived in either New South Wales (around 154,000) or Queensland (around 146,000), there was a disproportionately high number of observations from Victoria and the Northern Territory with 2,245 and 2,267 observations, respectively. The high number of observations in the Northern Territory is due to the fact that Indigenous people in the Northern Territory tend to be in smaller and more heterogeneous communities requiring a greater sample from each community to derive accurate results for each area (ABS 2008).

The main geographic record in the NATSISS is the State and Territory with the Australian Capital Territory (ACT) and Tasmania amalgamated. However, given around 70 per cent of Indigenous people lived outside the capital cities, we analysed the areas outside the capital cities based on the ABS remoteness structure which was available on the NATSISS. The remoteness structure is developed based on the road distance of an area to several closest service centres as indicated by population. With the remoteness structure, the ABS has created 17 areas with Queensland divided into four defined areas (Major Cities, Inner Regional, Outer Regional and Remote/Very Remote areas), New South Wales into three defined areas (Major Cities, Inner Regional, and Outer Regional), Victoria into two (Major Cities and Inner/Outer Regional) and Western Australia into two areas (Non-Remote and Remote/Very Remote). There are specific areas for the non-remote areas of South Australia and Tasmania as well as the remote area of the Northern Territory while the other three states/territories and the Australian Capital Territory are grouped as part of Balance of Australia.

Given the sensitivity of the data, the ABS has introduced a special arrangement for the dissemination of this survey. While the summary results from the 2008 NATSISS are available in a national level publication (see ABS 2010), the unit record data of NATSISS is not available to researchers. However, it is possible to have authorised access to the Confidentialised Unit Record File (CURF) through the ABS online query system by submitting code to the ABS Remote Access Data Laboratory (RADL) (ABS, 2006). The ABS can reject a request if it results in cells with a low number of observations or when the statistical or econometrical procedure uses a very specific sub-sample of the survey.

2.2 Methodology

Spatial microsimulation has emerged as a well-established technique in the estimation of small area statistics. Spatial microsimulation uses the individual or household units from survey data to populate each small area, subject to constraints from Census tables. These Census tables have a number of different classes in each table (for instance, there will be an estimate for the number of males aged 25 – 64 who are unemployed in a particular geographic area). These 'benchmark classes' in each benchmark table give information about the distribution at the smallest spatial area available in the data (Williamson et al., 1998; Voas and Williamson, 2000; Williamson, 2001).

Reweighting techniques have become one of the most common approaches to the creation of synthetic spatial microdata and, within this broad method, there are a number of different methodologies available (Tanton and Edwards, 2013; Anderson, 2007; Ballas, et al., 2005; Hynes, et al., 2009; Tanton et al., 2011; van Leeuwen et al., 2009; Voas and Williamson, 2000, Zaidi et al., 2009, Rahman et al., 2010).

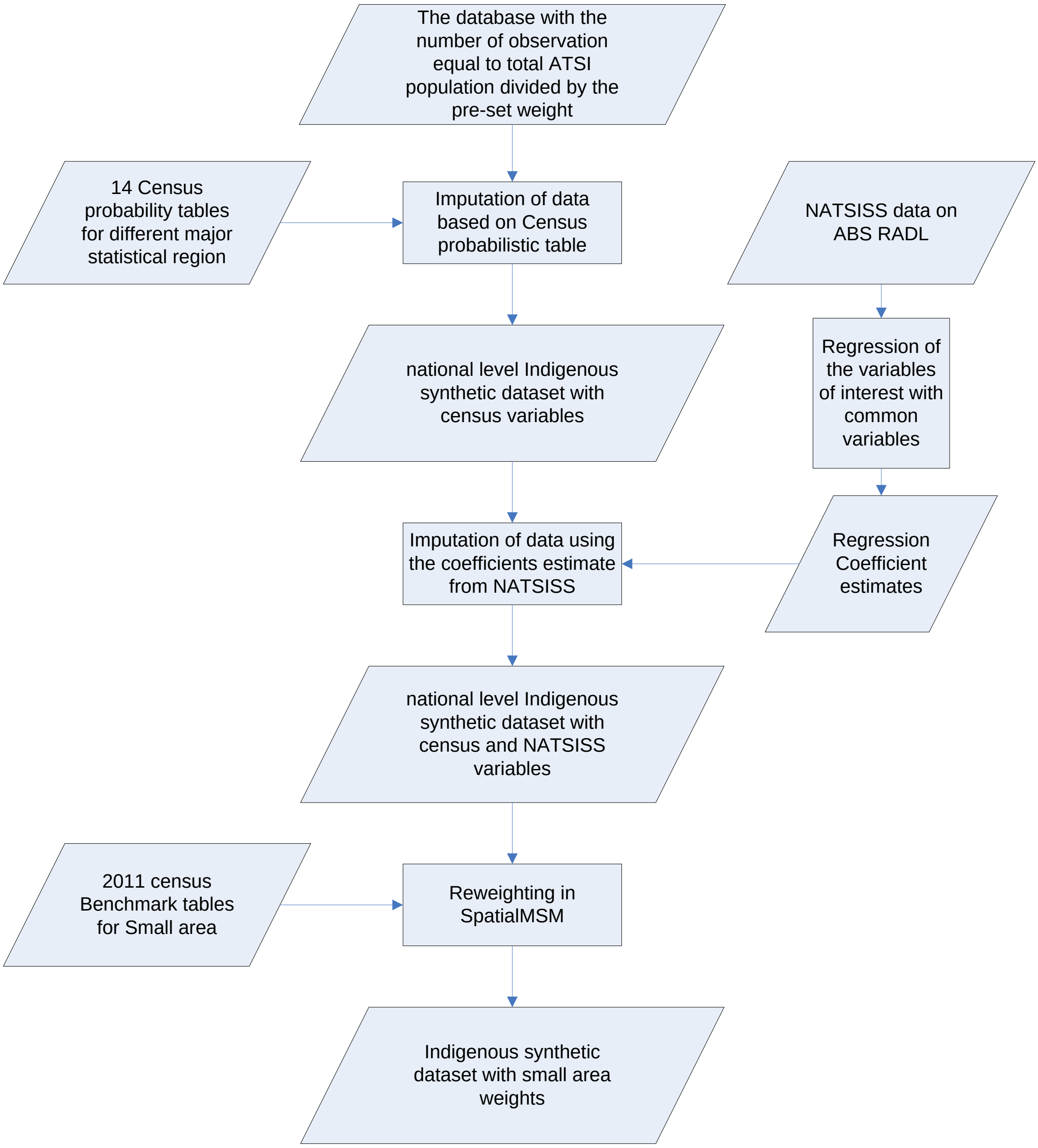
The SpatialMSM model used for this paper employs a generalised regression reweighting program from the Australian Bureau of Statistics (ABS) called GREGWT. The GREGWT algorithm uses a generalised regression technique to estimate weights for a household or individual from the survey, and then iterates until the weighted aggregate of the survey data produces characteristics that closely resemble the constraints for each small area (Bell, 2000; Tanton et al., 2011). The procedure can be classified as a deterministic method using formulae, similar to the iterative proportional fitting used by Anderson (2007) and Ballas et al. (2005), as opposed to a probabilistic method that pseudo-randomly selects households to fill an area described in other models (Voas and Williamson, 2000; Williamson et al., 1998). Despite the differences in approach, Tanton et al. (2007) confirms that the results from different reweighting methods are generally similar.

In Australia, spatial microsimulation has used unit record files from a number of surveys including the Survey of Income and Housing (SIH), the Household Expenditure Survey (HES) and The Household, Income and Labour Dynamics in Australia (HILDA). Given the unit record data from NATSISS survey was not available for this study, another set of synthetic data was created using two stages of an imputation method.

Figure 1 shows the flow chart of the estimation process including the two stages of imputation used to create the synthetic database. In the first stage, we imputed the data using a probabilistic table from Census data as in Williamson (2013). The main reason for using the probabilistic table was that the Tablebuilder program used to extract the Census data from the ABS allows us to create cross tabulations that can contain up to around 5,000,000 cells. This means we may create a set of probability tables with around four conditional factors.

As shown in Figure 1, this imputation would create a synthetic database made up of census data only. However, we also need the synthetic database to contain the variables that we would like to estimate from NATSISS. The probabilistic table technique cannot be used for this as the construction of the table will require the use of a relatively small number of observations to fill the cells of a cross tabulation, even with a small number of conditioning factors. Therefore, the second stage of the imputation process uses a regression method to impute the specific conditions that are available from the NATSISS detailed State by Remoteness (ABS 2008) table onto the synthetic database. Regression on variables of interest from the NATSISS will produce the coefficients needed to impute the variables onto the synthetic dataset. Given most of the variables of interest are likely to be binomial, the variables are estimated using a logit or probit regression model and the application of the coefficient will allow us to find the probability of the condition for each observation. A random number can then be applied to estimate the binomial value. Once all the necessary variables from NATSISS were successfully imputed, we can begin the reweighting process.

Figure 1 Flowchart of the estimation procedure


The database with the number of observation equal to total ATSI population divided by the pre-set weight
14 Census probability tables for different major statistical region
Imputation of data based on Census probabilistic table
NATSISS data on ABS RADL
national level Indigenous synthetic dataset with census variables
Regression of the variables of interest with common variables
Imputation of data using the coefficients estimate from NATSISS
Regression Coefficient estimates
national level Indigenous synthetic dataset with census and NATSISS variables
2011 census Benchmark tables for Small area
Reweighting in SpatialMSM
Indigenous synthetic dataset with small area weights

Note that there are some areas where an estimate cannot be produced by this reweighting process. This is mostly because the process does not achieve an acceptable error for the estimate. The error in the reweighting process is measured by the total absolute error (TAE) from all the benchmarks. The TAE is calculated by summing all the differences between the estimated number from the model and the benchmark number from every benchmark class of each benchmark table. The total error threshold that is set for this spatial microsimulation model is whether the TAE from all the benchmarks is greater than the population in that area. The TAE has been used in a number of spatial microsimulation models as a criterion for reweighting accuracy (Anderson, 2007; Williamson, et al., 1998) and has been assessed and supported by other studies such as Smith et al. (2009) and Voas and Williamson (2000).

2.3 The variables

This project aims to develop specific measures of wellbeing for the Indigenous population. The indicators chosen after some consultation with experts on Indigenous wellbeing are:

- participation in cultural activities;
- social capital;
- discrimination;
- health status;
- psychological stress;
- social and emotional wellbeing;
- financial stress;
- feelings of safety or stress;
- identification with clan, tribal or language group.

As highlighted in Cassells et al., (2010), the spatial microsimulation process starts by identifying relevant variables that exist in both databases. The identification of these variables in this study is not only important for the reweighting process but also for the imputation of data through both the probabilistic table and the regression process. The accuracy of spatial microsimulation can only be ensured when benchmarking variables are correlated with the variables being estimated from the model (Anderson, 2007).

2.3.1 The variables of interest

(a) Participation in selected cultural activities in last 12 months

This variable is available on the NATSISS as WCULACT and available for persons aged 15 years and above. This binary variable relates to the specific activities that have a cultural association including fishing, hunting, gathering wild plants / berries, making Aboriginal and Torres Strait Island arts or crafts, performing any Aboriginal and Torres Strait Island music, dance, theatre or writing or telling any Aboriginal and Torres Strait Island stories.

(b) Social capital

This variable is an index from -1 to 1 that was derived using Principal Component Analysis from several social capital variables including social cultural aspects such as involvement and attendance in cultural events, ceremonies or organisations (WCULEVNT andCULPQ12) and in sporting, social or community activities (WPAR3M). Social capital also considers the relationship to family and friends such as the frequency of contact with family or friends outside the household (FACCONC , NFACCOC and FANYCOC), whether the person can confide in family and friends(WFRNDCON) as well as the proportion of friends of the same age, Indigenous origin or similar education (PFRNDAGE, PFRNDIND, PFRNDEDU). Social capital also includes the ability to get or to give support outside the household (FCSPQ1, WSPTREL, WSPTANY), social efficacy (comfortable contributing in the family and community) and trust (in general and to the medical profession or police).

(c)	Felt discriminated against in last 12 months

This binary variable WDISC12 is the identification of discrimination in general. This includes discrimination at the workplace, at the neighbourhood and at educational institutions as well as while doing any sporting, recreational or leisure activities. The discrimination can also relate to treatment by the police, security guards, lawyers, in a court of law, by medical apparatus, or by any staff of Government agencies when seeking any public services.

(d)	Poor health status

This variable is based on self-assessed health (SAHQ1) where the person was asked to identify his or her health using the following categories: excellent, very good, good, fair and poor. There were more than 75 per cent of respondents rated their health as good or better, and around 10 per cent reported they had poor health. For this variable, we have used a binary response of feeling in poor health or not.

(e)	Psychological stress

This variable uses the categories 12 to 25 from the Kessler (K5) score that indicates whether a person has high or very high psychological distress. This measure is based on the psychological assessment questionnaire provided to the NATSISS surveyor to monitor non-specific psychological stress of the surveyed person. In Australia, the measure has commonly been used in public mental health services and more recently in various health surveys (Sunderland et al 2011).

(f)	Positive social and emotional wellbeing

This is one of only two continuous variables being estimated. This variable combines different indicators based on how the surveyed person felt in the last 4 weeks. This includes the feeling of calm and peaceful (SEWQ12), happy (SEWQ13), full of life (SEWQ14) and having a lot of energy (SEWQ15).

(g)	Household members ran out of money for basic living expenses

This binary variable indicates whether household members ran out of money for basic living expenses. The question was asked twice in the questionnaire, first based on the condition in last 12 months(HHFSQ4) and second based on condition in last 2 weeks (NOMON2W).

(h)	Feelings of safety walking alone in the local area after dark

Unlike the other variables, this variable is estimated using two steps. We first flag those who are walking alone after dark, and then look at who felt safe among those who were walking alone after dark. The original variable measuring feelings of safety actually contains five values: very safe, safe, neither safe nor unsafe, unsafe and very unsafe. The unsafe and very unsafe were combined to get those who were not feeling safe walking alone after dark.

(i) Whether victim of physical/threatened violence

The variable is a combination of two variables on the NATSISS with a value of one assigned to those who have either been a victim of physical violence (CJVQ1) or were threatened with physical violence in last 12 months.

(j) Personally experienced stressor

This variable identifies those who are experiencing different stressors in last 12 months (WSTROWN). There are around 24 types of stressor used for this variable including bad illness, bad accident, marriage, pregnancy, divorce or separation, death of family member or close friend, lost or change job, alcohol or drug related problems, abuse or violent crime and discrimination.

(k) Identifies with clan, tribal or language group

For this variable, the reference person in the family is asked whether he or she and the family identify themselves into a certain tribal group, language group, clan, mission or regional group.

2.3.2 Benchmarking variables between the Census and NATSISS

The following are the variables that are not only available in both Census and NATSISS but also defined in a similar fashion so they can be benchmarked.

(a) Age of person

This variable in the Census indicates the age of a person based on the last birthday before the August Census collection date. The date of birth information provides the best estimate of age if it is available. Alternatively, stated age will be used only if date of birth is not provided. In the NATSISS we use AGEC. This variable is based on the age stated by the reference person. Similar to the age in Census, the age variable in NATSISS is mostly in single years. However, on the NATSISS the ABS confidentialises those who are older than 64 years old. We have therefore had to aggregate the age ranges to match between the two sources.

(b) Sex of a person

SEXP identifies the sex of a person. In both Census and NATSISS, there are only two options in this category, male or female, and if there is no answer on the NATSISS then the sex variable is imputed.

(c) Relationship in household

The relationship of a particular person within the household is another variable available on both the Census and the NATSISS. On the Census, the variable (RLHP) contains 29 different relationship types including family and non-family member with categories such as Husband, Wife in a registered marriage, Partner in a de facto marriage, Lone parent, Natural or adopted child, Step child, and so on. The variable (RLHHLC) on the NATSISS contains only eight groups husband, wife or partner; lone parent; child under 15; dependent student; non-dependent child; other related individual; non-family member and lone person. We have therefore aggregated the Census classifications to match the survey classifications.

(d) Level of highest non-school qualification

The Non-School Qualification on the census (QALLP) describes the level of education of the highest completed qualification after secondary school. Although the definition of the matching variable in the NATSSIS is practically the same, the classification is slightly different and regrouping on both databases was needed to produce the same classification across these two databases. The not applicable category in both databases includes people with no qualification, people still studying for a first qualification and people aged under 15 years.

(e) School status

This variable indicates whether a person is attending school on a full time or part time basis. The variable on the Census (STUP) is not exactly the same as the matching variable on the NATSISS (FTPTSTDY) as the variable on the NATSISS is only applied to those aged 15 years and over while in the census, STUP is applied to everyone. Therefore, we have adjusted the Census variable to represent only those who are 15 years and above.

(f) Labour force status

Although the two datasets use a slightly different definition, the difference between the labour force status classification on the Census and the NATSISS is not great. On the Census, the variable LFSP indicates the person's labour force status in the week before the Census night, while on the NATSISS (EMPSTAC) the reference period was the entire week, Monday through Sunday, prior to the interview.

(g) Occupation in main job

In both the Census and the NATSISS, the occupation variable shows the occupation of any employed person classified using the new Australian New Zealand Standard Classification of Occupations (ANZSCO). The classification includes Managers, Professionals, Technicians and Trades Workers, Community and Personal Service Workers, Clerical and Administrative Workers, Sales Workers, Machinery Operators and Drivers and Labourers.

(h) Equivalised household gross weekly income

Income plays in important role in wellbeing. The variable Total Household Income (weekly) in the Census data (HIND) is calculated by summing the personal incomes reported by all household members aged 15 years and over. The Census collects personal income in ranges. Therefore, a specific dollar amount needs to be allocated to each person using median incomes for each range based on data from the Survey of Income and Housing. Household income is not calculated where a household member aged 15 years and over did not state their income, or was temporarily absent. These households are coded to the 'Partial income stated' category. The definition of the Household gross weekly income variable (WINCPH) in NATSISS is similar but it is presented as single dollar value rather than in income groups.

This study further calculates equivalised household gross weekly income by adjusting the household income by the number of people in the household. This adjustment is necessary to compare the incomes of households with a different number of people in them. The adjustment uses the 'modified OECD' equivalence scale, which is built up by allocating 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 (Atkinson et al, 1995).

(i) Family type

This variable classifies families into different types of families. There are different categorisations of family type on the Census. The main category only contains four types of family: couple without children, couple with children, lone parent and other family while the most detailed breakdown has 17 categories. Family type on the NATSISS contains 13 categories that do not exactly match the categories in the Census. Therefore, the family type in the NATSISS was grouped to match the more general four categories of family type in the Census.

(j) Tenure type and Landlord type

The variable for tenure type is TEND in the Census and HSTENUC in the NATSISS. The variable describes whether a dwelling is owned, being purchased or rented. The landlord type variable (LLDD on the Census and TYPRENC on the NATSISS) records the landlord type of rented dwellings. Although not exactly the same, the categories from each data source are similar. For example, the dwelling categorised as being purchased on the Census is comparable with owner with a mortgage on the NATSSIS. This is also the case with landlord type. Moreover, the landlord type is aggregated further to be public (rented from State or Territory housing authority), private or other.

(k) Monthly mortgage repayments

On the 2011 Census, the value of mortgage repayments only is available (MRED), which matches the NATSISS definition. One difference is that the NATSISS variable is calculated on a weekly basis while the Census data is monthly. We have converted the weekly amount to monthly amounts.

(l) Working motor vehicles owned by household members

There is a slight difference in definitions in the motor vehicle benchmark. The variable on the Census (VEHD) refers to the number of registered motor vehicles owned or used by household members garaged or parked at or near a private dwelling on Census Night. On the NATSISS, the variable (NCARHHC) is defined as the number of working motor vehicles owned by household members.

(m) Computers connected to the Internet in the household

The variable on the NATSISS (HHITHQ6) asks whether any computers are connected to the Internet in the household, while the variable on the Census (NEDD) records the type of Internet connection most frequently used in addition to whether the dwelling has an Internet connection. For this model we have just used whether the household has or doesn't have an internet connection.

3 Building the Synthetic Unit Record data

3.1 Creation of the unit record data based on Census tables

For spatial microsimulation, a survey dataset is required that provides the unit records. Traditionally, we would use ABS Confidentialised Unit Record Files for this, but there is not one available for the NATSISS, so we had to create one. The first step in the construction of the database was to create some empty observations. The empty observations contain only three variables: an observation number or person identifier, the geographic area from the Census and an initial weight. To get this weight, we need to know how many Indigenous people live in different areas of Australia. Therefore, the first thing we needed to do was decide which of the ASGC areas to use for this part of the analysis. This choice of area will affect the accuracy of our final reweighting, as knowing approximately where someone lives gives a starting point for the estimation process.

The geographic area chosen for the first part of this estimation procedure was the Major Statistical Region from the ASGC structure. There are 14 regions in this classification: the Capital City and the balance of the State of the five larger States, namely, New South Wales, Victoria, Queensland, South Australia and Western Australia and one area each for Tasmania, Northern Territory, the Australian Capital Territory and Other Territories. In this project, we excluded the “Other territory” region which has a very small population.

Based on the number of Indigenous people in the area from the Census, each region was populated with synthetic people. Note that this does not mean that the number of synthetic people in a region will match the actual number of people in that region as the number of observations in places like Sydney would be too large for the model to work efficiently. In creating the synthetic people, each synthetic person receives a weight to represent a number of people in the area. The initial weight is calculated as the actual number of people in the area from the Census divided by the number of synthetic people.

After this synthetic database was created, six variables were imputed to each of the observations. These six variables were age, sex, labour force status, gross weekly household income, gross weekly household equivalised income, and family composition. For this project, age was reclassified into eight age groups: 0 to 14, 15 to 24, 25 to 34, 35 to 54, 55 to 64, 65 to 74, 75 to 84, and 85 years and over; and the labour force status has been simplified into employed, unemployed, not in the labour force, not applicable, and not stated. Not applicable applies to people who are not old enough to be employed (so aged under 15). For family composition, the classifications were couple family with no children, couple family with children, one parent family and other family. The not applicable classification was also used for those living in lone person households and group households.

To allocate these six variables to the records in our synthetic database, we have used a probability table from a cross tabulation of these six variables from the Census data. Overall, there were 126 thousand categories allocated to our population. The use of only six variables for the first imputation was based on limitations of the Census Tablebuilder program.

The next variable imputed was Individual gross weekly Income. This variable shows the income level of people aged 15 years and over in income ranges. There were 15 different categories with 10 valid income brackets and the categories negative income, nil income, not stated, not applicable and overseas visitor.

To impute the Individual gross weekly income bracket for each person we used the probability of the person being in a particular income group based on five other variables that have been imputed previously – age, sex, labour force status, gross weekly household income and gross weekly household equivalised income.

We then imputed two variables about the number of people in the family – the number of dependent children in the family and the number of people in the family.

The variable for the number of dependent children in the family gives the count of children under 15 years of age, or dependent students aged 15-24 years, in the family. Although this can include up to three dependent children who were temporarily absent on Census night, the variable itself has a limitation. It provides an accurate count for up to five children and is then top coded, so the final category is six or more children.

Unlike the number of dependent children in the family, the number of people in the family includes other related individuals who are not part of the nuclear family, such as in-laws, grandparents, uncles and so on.

These two variables were imputed together because we could get more accurate estimates using a combination of equivalised income and non-equivalised income as the main variables in the probability table. Besides those two variables, we also used sex and family composition as the basis for the joint probability.

After imputing these variables, we then imputed the tenure type. This variable indicates whether the dwelling the household lives in is a rental property or owned. This variable is only available for occupied private dwellings. The classification used for this variable is fully owned, being purchased, rented, other tenure, not stated, and not applicable.

To impute this variable we used a probability table based on five variables – age, sex, family composition, gross weekly household income and gross weekly household equivalised income. In estimating this variable, we started to find that there were some observations that did not fit into the characteristics shown in the probability table, so we decided to drop these observations from our synthetic dataset. The reason these observations may occur is that the imputation may assign a certain condition to an observation where, in the census, the condition does not exist in the area. For example, the modelling may impute a family with children with an income of $100 - $200 per week when no family with children earning this much exists in the area.

We then estimated the number of people usually resident in the dwelling. Unfortunately, due to limitations in Table Builder, we had to drop the tenure variable and use the number of dependent children in the family, the number of people in the family, household equivalised income and household income in the probability table. All these variables combined should give a sense of how many people actually live in the household as the equalivisation of Household Income is based on the number of adults and children in the household.

The next indicator imputed consists of two variables combined together – the highest year of school completed, and the highest completed non-school qualification. These combine to form a level of education. The highest year of schooling completed shows the highest level of primary or secondary schooling completed and has categories of year 12 or equivalent, year 11 or equivalent, year 10 or equivalent, year 9 or equivalent, year 8 or below and did not go to school. People aged below 15 years are not applicable for this category. The highest completed non-school qualification includes postgraduate degree level, graduate diploma and graduate certificate level, bachelor degree level, advanced diploma and diploma level as well as certificate level.

These two variables are strongly correlated to each other since only a very few people have a Bachelor degree and above qualification without having finished year 12. Personal income, sex, age and labour force status are variables that are closely correlated with these education variables, and hence are used in the probability table.

The next variable imputed was the relationship in household. This variable describes the relationship of each person in the household to the household reference person. This includes family and non-family member with categories such as husband, wife in a registered marriage, partner in a de facto marriage, lone parent, natural or adopted child, step child, foster child, grandchild, unrelated child, brother/sister, father/mother, cousin, uncle/aunt, nephew/niece and so on.

The variables used to construct the probability table for this variable were age, sex, family composition, labour force status and personal income.

After imputing the relationship variable, we imputed a combined variable which consisted of landlord type and weekly rent. These two variables give a specific attribute to those living in a rented dwelling. The landlord type variable shows the landlord type of the rented property categorised as private tenure, State or Territory housing authority (public) and other tenure. The rent is divided into 16 brackets from zero to $49 a week to 550 a week and over.

The variables used to distribute the probability of renting are the number of people usually resident in the dwelling, the number of dependent children in the family, tenure type and gross weekly household income.

The next variable imputed was the monthly housing loan repayment. This variable was only attributed to those households who were purchasing their dwelling. There are 18 payment brackets associated with this variable, and the probability has been estimated based on the distribution of the number of people usually resident in the dwelling, the number of dependent children in the family, age, tenure type and gross weekly household income.

The next two variables imputed were dwelling related variables, being the type of internet connection and the number of motor vehicles. The type of internet connection variable shows the most frequently used type of Internet connection (separating broadband, dial-up and no connection) in a dwelling. The variables used for the probability table for the imputation of this variable were number of people usually resident in the dwelling, the count of dependent children in the family, age, tenure type and gross weekly household income.

The number of motor vehicles variable shows the number of registered motor vehicles, including company owned vehicles, owned or used by household members. This variable is only applicable to occupied private dwellings. The variables used for the probability table to impute this variable were number of people usually resident in the dwelling, the number of dependent children in the family, age, tenure type and gross weekly household income.

We then imputed two variables related to employment: occupation and industry of employment. The variables used to distribute the probability of each person’s occupation were age, sex, qualifications, labour force status and individual gross weekly income.

The industry of employment variable shows the Industry that a person is employed in based on the Australian and New Zealand Standard Industrial Classification (ANZSIC) 2006 classification. The ABS classifies a person’s industry of employment based on the description of the business, and the main goods produced, or main services provided in the place the person works. In addition, the name of the business, the employed person's occupation and main tasks and duties may also be used as additional information to determine the industry. As the occupation of the person determines the industry of employment, occupation is one of the variables used for the industry of employment probability matrix. Other variables used were age, sex, qualifications and individual gross weekly Income.

In addition to the variables above, we also imputed the full time/part time student status and the usual address on census night indicator.

This Full Time/Part Time student status variable is imputed because we wanted information on the school retention rate among Indigenous students. The variables used to construct the probability table for this variable were age, sex, family composition, labour force status and household equivalised income.

The usual address on Census night indicator shows whether the person actually lives in the dwelling on a daily basis. This variable is needed because we have based our imputation on enumerated persons, so this variable identifies those living at their usual address. The imputation used age, sex, labour force status, personal income and the relationship in household in the probability tables.

Table 1 summarises this imputation process, providing a summary of all the variables imputed and the variables used for the conditional probabilities.

Table 2 The sequences of imputation and the variables used for the conditional probability

imputation step	Variables Imputed	Conditional Probability Basis
1	age, sex, labour force status, gross weekly household income, gross weekly household equivalised income, and family composition	age, sex, labour force status, gross weekly household income, gross weekly household equivalised income, and family composition
2	individual gross weekly income	age, sex, labour force status, gross weekly household income and gross weekly household equivalised income
3	the number of dependent children in the family and the number of people in the family	equivalised income, non equivalised income, sex and family composition
4	tenure type	age, sex, family composition, gross weekly household income and gross weekly household equivalised income
5	number of people usually resident in the dwelling	Number of dependent children in the family, number of people in the family, household equivalised income and household income
6	the highest year of school completed, and the non-school qualification: level of education	personal income, sex, age and labour force status
7	relationship in household	age, sex, family composition, labour force status and personal income
8	landlord type and weekly rent	number of people usually resident in the dwelling, number of dependent children in the family, tenure type and gross weekly household income
9	monthly housing loan repayment	number of people usually resident in the dwelling, number of dependent children in the family, age, tenure type and gross weekly household income
10	type of internet connection	number of people usually resident in the dwelling, number of dependent children in the family, age, tenure type and gross weekly household income
11	number of motor vehicles	number of people usually resident in the dwelling, number of dependent children in the family, age, tenure type and gross weekly household income
12	occupation	age, sex, qualifications, labour force status and individual gross weekly income
13	industry of employment	occupation, age, sex, qualifications and individual gross weekly income
14	full-time/part-time student status	age, sex, family composition, labour force status and household equivalised income

The synthetic database that we have created using this process may not produce a set of observations that exactly match the Census cross-tabulations. The next step was to apply a reweighting method to make the current synthetic database much closer to the original Census tables. Before doing this, we decided to look at the differences between tables calculated from the synthetic dataset and census data for several of these variables before the reweighting process took place to analyse how well the synthetic dataset is able to replicate the actual census profile. The figures shown in Table 2 are the total absolute error calculated as the sum of the absolute differences between the Census estimate and Synthetic estimate divided by the population of the area - see Section 2.2 above.

As can be seen in Table 2, the differences are around 10 per cent for all variables except education, where the difference is around 20 per cent. The ACT also has greater differences, and this may be because the number of Indigenous people in Canberra is very small. The education variable and number of children are the variables with the lowest level of accuracy.

To get better estimates, we then used a reweighting program to benchmark the synthetic dataset to Census data at the Capital City/Balance of State level. The variables benchmarked for this reweighting process were age, sex, income, number of children, education, tenure, internet, vehicle ownership and occupation. Although not shown in Table 2, the reweighting resulted in a total error of around three per cent.

Table 2 Differences (TAE) between Census proportions and synthetic proportions for a number of variables by Section of State

Variable	Sydney	NSW BS	Melbourne	Victoria BS	Brisbane	Queensland BS	Adelaide	SA BS	Perth	WA BS	Tasmania	NT	ACT
CDCF Count of Dependent Children in Family	0.144	0.156	0.218	0.195	0.161	0.123	0.214	0.165	0.233	0.138	0.210	0.095	0.269
QLLP and HSCP (Education)	0.195	0.208	0.220	0.234	0.181	0.216	0.226	0.245	0.103	0.121	0.148	0.128	0.337
HLRD Housing Loan Repayment (monthly)	0.039	0.023	0.079	0.049	0.035	0.020	0.065	0.069	0.053	0.031	0.059	0.021	0.046
INCP Individual Income (gross weekly)	0.075	0.052	0.063	0.048	0.076	0.068	0.102	0.059	0.040	0.078	0.053	0.052	0.193
NEDD Type of Internet Connection	0.053	0.065	0.104	0.080	0.055	0.064	0.131	0.138	0.110	0.096	0.054	0.089	0.061
NPRD Number of Persons Usually Resident in Dwelling	0.074	0.069	0.118	0.093	0.080	0.040	0.073	0.103	0.176	0.147	0.157	0.060	0.141
OCC06P Occupation 06 (ANZSCO)	0.041	0.034	0.050	0.034	0.054	0.047	0.077	0.056	0.033	0.052	0.029	0.036	0.192
RLHP Relationship in Household	0.089	0.074	0.148	0.168	0.105	0.061	0.188	0.168	0.172	0.108	0.125	0.078	0.327
STUP Full/Part-Time Student Status	0.002	0.011	0.015	0.058	0.029	0.019	0.095	0.083	0.032	0.013	0.016	0.030	0.103
TEND Tenure Type	0.118	0.074	0.138	0.138	0.130	0.107	0.157	0.233	0.149	0.200	0.121	0.158	0.172
UAICP Usual Address Indicator Census Night	0.020	0.015	0.010	0.016	0.023	0.011	0.013	0.017	0.038	0.042	0.005	0.009	0.048
VEHD Number of Motor Vehicles	0.086	0.091	0.122	0.090	0.085	0.067	0.177	0.157	0.146	0.144	0.099	0.099	0.140
AVERAGE	0.078	0.073	0.107	0.100	0.084	0.070	0.126	0.124	0.107	0.097	0.090	0.071	0.169

Source: ABS Census and NATSEM synthetic data

3.2 Imputing the NATSISS variables

The next step was to impute the specific conditions that are available from the NATSISS 2008 Detailed State by Remoteness table onto the synthetic database. This imputation is performed in two or three stages. The first stage involves the regression of the variables of interest from the NATSISS to obtain the coefficients needed for the imputation of the variables onto the synthetic dataset. The second stage involves using these coefficients to impute the new variables onto the synthetic dataset. The third stage is needed if the variable is estimated using a logit or probit regression model and a random number needs to be used to estimate the binomial value using the probability from the regression.

Before conducting these stages, we need to set up the variables in NATSISS so the coefficients produced by the regression can be used in the synthetic database. Given the variables in this database are all categorical, including the income as it is in ranges, the regression also needs to use categorical variables. Therefore, dummy variables were created in the NATSISS to be the independent variables in the regression. The dummy variables reflect the categories used in the synthetic database, so the variables on the NATSSIS have to be reclassified to ensure that they have the same categories as the synthetic database.

3.2.1 Independent variables for the regression

This section outlines the creation of the independent variables.

(a) Age of person

There are three dummy variables for age:

- People aged 25 to 34
- People aged 35 to 54
- People aged 55 and older

The base value is those younger than 25 years old. Depending on the variable being estimated this means 0-24 or 15-24 years old.

(b) Sex of a person

This variable uses the base value as male.

(c) Relationship in household

The relationship variable uses binary dummy variables:

- Husband, wife or partner
- Lone Parent
- Other relative such as uncle, grandparent etc
- Not family related
- Lone person

The base value is children and dependent students.

(d) Level of highest non-school qualification

The binary dummy variables used are:

- Those with Bachelor degree and above including Master and PhD
- Those with diploma

- Those with certificate

The base value is those without a qualification. This includes children over 15 who are still in school.

(e) School status

There are binary variables for two types of students:

- Full time student age 15 years and over
- Part time student age 15 years and over

The base value is those who are not in the education system.

(f) Labour force status

There are two dummy variables used:

- Employed
- Unemployed

The base value is those who are not in labour force.

(g) Occupation in main job

Only two occupations have dummy variables:

- Those who work as managers and professional
- Those who work in other occupations

The base value is anyone who is not working including unemployed and those who are not in the labour force.

(h) Equivalised household gross weekly income

Dummy variables are used for:

- Those in a high income family with equivalised income more than or equal to $1300/week
- Those in a medium income family with equivalised income more than or equal to $600/week but less than $1300/week
- Those in a low income family with equivalised income more than or equal to $250/week but less than $600/week

The base value is those living in a household with equivalised household gross weekly income of less than $250/week.

(i) Family type

Three dummy variables are used:

- Couple without children
- Couple with children
- Single Parent

The base value is lone person and group household together.

(j) Number of children in the family

The number of children in the family is from none to five or more.

(k) Tenure type and Landlord type

For tenure type, there are dummy variables for:

- Those who already fully own their house
- Those who are still paying mortgage of their house

This variable is combined with the rental variable as follows:

- Those who are privately renting;
- Those who are renting from a government authority
- Those who are renting from another entity or have another rental arrangement

(l) Working motor vehicles owned by household members

This variable has a dummy variable for:

- Those who do not have a motor vehicle at their dwelling

The base value is that there is a motor vehicle at the dwelling.

(m) Computers connected to the Internet in household

This dummy variable is for:

- Those with an internet connection in the dwelling

The base value is no internet connection.

3.2.2 Regression statistics

Given that most of the variables are binary variables, we mostly used a probit regression model. The model was run separately for different areas as the log-likelihood test for the probit regression indicated that the area regressions provides significantly better goodness of fit (ie, the log-likelihood for the Australia wide model was much lower than the log-likelihood for the area based models).

The problem applying the regression on an area basis was that the areas in the NATSISS and in the synthetic database were not perfectly compatible. As mentioned in section two, the most detailed breakdown in NATSISS is the 17 areas based on the ABS remoteness area. Using this geography, it was possible to get coefficients for Sydney, Melbourne, Brisbane and Perth from the major cities of New South Wales (NSW), Victoria, Queensland and Western Australia, respectively. To estimate Adelaide, we used the South Australia Non-remote area while the South Australia Balance of State area uses data from Balance of Australia - Remote/Very Remote area. Two areas in the synthetic data – Tasmania and the Northern Territory (NT) – use different combinations of data for the regression. Tasmania uses “Tasmania non-Remote” and “Balance of Australia - Remote/Very Remote” and the NT uses “Remote/Very Remote” and “Balance of Australia - Non-remote”.

Tables 3a and 3b show the results from the probit and linear regressions, respectively, used to estimate each variable of interest from the NATSISS. The variable names after each area are the independent variables that were most significant in explaining the dependent variable. It can be seen that these are different for each area.

Pseudo-R^2 and log-likelihood estimates are given for the binary variables where a probit regression was used, and R^2 is given for the continuous variables (Social Capital and Positive Social and Emotional Wellbeing). The main determinant is the most significant independent variables used for each area in the regression.

The much higher log-likelihood values in Table 3 indicate that the regressions using the above areas do provide better results for most areas and variables. This may be because the values will be different in each area, so estimating each area separately should give better results; but it may also be because a particular dependent variable in the regression could have different determinants in different areas, so the independent variables may have greater explanatory power. It is important to note that although the results from the area based analysis are better, only a few of the results show an excellent goodness of fit (Pseudo-R^2).

Table 3a Probit regression statistics and determinants for each variable of interest

Variable	Main determinant	Pseudo-R^2	Log likelihood
Participation in selected cultural activities	The entire Australia regression	8.01	-4548.90
	Sydney: tenure	9.91	-242.04
	NSW-BS: area, tenure	10.34	-333.18
	Melbourne: family type	5.85	-452.86
	Victoria-BS: area, tenure	12.78	-320.27
	Brisbane: family type	24.95	-80.10
	Queensland-BS: family type, education	13.92	-496.77
	Adelaide: tenure	10.04	-347.93
	SA-BS: family type	20.10	-186.00
	Perth: tenure	13.81	-165.46
	WA-BS: tenure, occupation	14.26	-384.35
	Tasmania: family type, language	14.30	-298.06
	NT: area, education	12.14	-382.51
	ACT: area, family type	11.69	-437.26
Felt discriminated against	The entire Australia regression	5.21	-4095.75
	Sydney: tenure, area	14.48	-177.24
	NSW-BS: language, education labour force	9.84	-287.37
	Melbourne: tenure, family type, education	12.33	-345.99
	Victoria-BS: language, family type, tenure	11.27	-294.06
	Brisbane: tenure, income, labour force	33.35	-63.64
	Queensland-BS: tenure, income	8.00	-521.24
	Adelaide: family type, income	10.37	-321.35
	SA-BS: education, income	5.63	-177.77
	Perth: tenure, area	13.87	-160.38
	WA-BS: area, occupation	4.81	-459.55
	Tasmania: area, tenure	8.02	-176.94
	NT: family type, area, education	10.22	-400.00
	ACT: family type, area	8.05	-494.59
Poor health status	The entire Australia regression	18.98	-1963.19
	Sydney: area, tenure	35.76	-108.49
	NSW-BS: family type	22.81	-146.19
	Melbourne: area, tenure	23.74	-187.63
	Victoria-BS: area, tenure	30.71	-118.25
	Brisbane: education	49.36	-18.81
	Queensland-BS: area, tenure	25.88	-178.07
	Adelaide: area, education	14.30	-179.17
	SA-BS: area, family type	25.12	-81.59
	Perth: family type	33.42	-58.80
	WA-BS: area, tenure	28.09	-147.78
	Tasmania: area, family type	26.85	-131.64
	NT: area, age, labour force	22.44	-178.51
	ACT: area, age, labour force	20.25	-216.44

Variable	Main determinant	Pseudo-R^2	Log likelihood
High Kessler (K5) score	The entire Australia regression	4.27	-4370.08
	Sydney: tenure, family type	19.32	-202.24
	NSW-BS: area, family type, tenure	9.53	-310.69
	Melbourne: family type, labour force, education	13.16	-385.13
	Victoria-BS: tenure, area	9.38	-323.13
	Brisbane: area, family type	10.93	-85.79
	Queensland-BS: income, area, family type	4.56	-555.56
	Adelaide: area, family type	6.31	-337.80
	SA-BS: family type, age, internet	8.46	-190.25
	Perth: area, tenure	13.19	-147.94
	WA-BS: area, study, internet	7.77	-443.07
	Tasmania: family type, age-sex, area education	9.66	-268.18
	NT: occupation, area, tenure	7.98	-404.03
	ACT: area, family type, occupation	3.14	-523.22
Household members ran out of money for basic living expenses	The entire Australia regression	6.91	-6537.20
	Sydney: family type	15.99	-336.29
	NSW-BS: area, family type	8.37	-496.45
	Melbourne: income, tenure, family type	13.46	-591.04
	Victoria-BS: family type, area	10.45	-509.34
	Brisbane: family type, area	20.82	-117.44
	Queensland-BS: area, tenure, age	6.62	-738.50
	Adelaide: income, labour force, family type	11.66	-467.57
	SA-BS: area, family type, tenure	20.26	-245.31
	Perth: area, tenure	17.05	-232.06
	WA-BS: tenure, income	10.39	-695.93
	Tasmania: area, tenure	17.40	-336.93
	NT: tenure, age	12.37	-634.37
	ACT: tenure, age, labour force	4.57	-898.40
Feelings of safety walking alone in local area after dark	The entire Australia regression	12.44	-3494.43
	Sydney: sex, family type, income	19.74	-182.84
	NSW-BS: area, family type, tenure	14.73	-237.96
	Melbourne: area, family type, tenure	17.63	-323.44
	Victoria-BS: sex, area, tenure, labour force, age	24.30	-225.12
	Brisbane: family type, sex	21.37	-65.17
	Queensland-BS: area, sex, family type	15.15	-453.40
	Adelaide: tenure	18.09	-274.72
	SA-BS: education, tenure, income	23.74	-103.08
	Perth: area, tenure	26.97	-118.09
	WA-BS: area, sex, study, family type	13.39	-333.39
	Tasmania: tenure, area	22.88	-167.70
	NT: area, sex, tenure	15.65	-340.42
	ACT: area, tenure, sex	10.72	-436.17

Variable	Main determinant	Pseudo-R^2	Log likelihood
Whether victim of physical/threatened	The entire Australia regression	6.48	-3885.40
	Sydney: tenure	12.16	-192.41
	NSW-BS: area, family type	9.42	-281.79
	Melbourne: area, family type, age	14.39	-346.22
	Victoria-BS: tenure, area	13.07	-287.90
	Brisbane: age, family type, area	14.97	-79.19
	Queensland-BS: family type, age, tenure, area	9.36	-426.52
	Adelaide: age, tenure, education	8.97	-300.12
	SA-BS: family type, area, tenure	21.20	-132.19
	Perth: area, tenure	18.49	-142.39
	WA-BS: age family type, tenure	6.81	-429.85
	Tasmania: tenure	15.94	-221.51
	NT: family type, age	12.13	-356.18
	ACT: family type	10.06	-431.43
Personally experienced stressor	The entire Australia regression	3.27	-4851.67
	Sydney: area, family type	8.53	-245.32
	NSW-BS: area, family type	8.96	-338.12
	Melbourne: area, tenure	8.59	-412.88
	Victoria-BS: area tenure	7.53	-320.33
	Brisbane: tenure, income, labour force	13.40	-93.74
	Queensland-BS: labour force, family type, education	3.79	-640.34
	Adelaide: area, education, occupation, labour force	6.20	-318.26
	SA-BS: income, study, tenure, language	13.73	-195.51
	Perth: tenure area	10.81	-162.56
	WA-BS: labour force, tenure, education	6.94	-478.90
	Tasmania: education, income, language	11.18	-304.04
	NT: area, labour force, study, tenure, internet	5.40	-476.88
	ACT: area, family type, labour force, age	4.67	-597.22
Identifies with clan, tribal or language group	The entire Australia regression	9.35	-7096.24
	Sydney: area, tenure	11.28	-390.78
	NSW-BS: area, education, family type	9.58	-532.10
	Melbourne: family type, tenure, education	6.98	-699.88
	Victoria-BS: area, tenure	7.55	-538.71
	Brisbane: family type, age	13.98	-142.69
	Queensland-BS: age, education, family type Adelaide:	7.62	-910.94
	area, tenure, family type	9.89	-494.06
	SA-BS: language, education, age	16.78	-290.47
	Perth: tenure, area	12.92	-267.91
	WA-BS: family type, occupation, area	11.88	-720.30
	Tasmania: language, tenure, age	11.74	-433.06
	NT: education, language occupation	8.94	-522.64
	ACT: education, language, area	8.54	-534.35

Table 3b Linear regression statistics and determinants for each variable of interest

Variable	Main determinant	R^2
Social capital	The entire Australia regression	7.07
	Sydney: family type, income	17.49
	NSW-BS: education, tenure, income	18.35
	Melbourne: area, language, occupation	17.09
	Victoria-BS: tenure, language	13.40
	Brisbane: education, language	30.31
	Queensland-BS: area, family type	13.46
	Adelaide: tenure, language, income	9.13
	SA-BS: family type, labour force	14.68
	Perth: language, education, occupation	19.92
	WA-BS: education, occupation, area	12.41
	Tasmania: tenure	15.20
	NT: education, tenure, area, labour force	13.88
	ACT: labour force, tenure education	10.55
Positive social and emotional wellbeing	The entire Australia regression	9.23
	Sydney: area, family type, labour force, tenure	14.53
	NSW-BS: area, family type	11.30
	Melbourne: area, language, labour force	13.47
	Victoria-BS: area, tenure	12.84
	Brisbane: area, age, family type, language	17.37
	Queensland-BS: area, income, tenure	8.90
	Adelaide: area, family type, tenure	12.39
	SA-BS: area, income, family type, age	20.24
	Perth: area, tenure	15.65
	WA-BS: area, family type, tenure, labour force	11.01
	Tasmania: area, labour force, income, family type, language	16.58
	NT: area, tenure, language	12.15
	ACT: area, tenure, family type	9.75

3.3 Overall imputation result

The next step was to compare the estimated values with the known values in the NATSISS. Although the regression statistics do not indicate strong predictive power for individual persons, it may still be possible to get a reasonable estimate for an overall area. The estimation is first conducted at a person level and then the person level estimates are aggregated to get an area level estimate. For the two continuous variables (social capital and positive wellbeing), the coefficient can be used to directly estimate the value of the variable while for the probit model, the regression coefficient produces the probability that the variable is one and the variable is estimated by randomly assigning a value of one for each observation based on this probability.

The assessment on how well the imputation worked is done by comparing estimates from the post-imputation synthetic database against estimates from the original NATSISS at the State/Territory level. Overall, the results indicate that we have a synthetic database than can provide good estimates. For this comparison, we compared proportions rather than absolute values as the base population for each dataset is different (2006 in our synthetic database based on 2006 Census data and 2008 in the NATSISS). Therefore, we used the large sample test for the population proportion for measuring the confidence level of our initial estimate. This test statistic can be represented by

$$t = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

where $\hat{p}$ is the estimated proportion and p_0 is the proportion of the variable of interest that came from the NATSISS. The value of n represents the number in the sample and we have used the sample in each area from the NATSISS for this number.

Table 4 shows the results from this statistical test. In the table, the difference is the percentage point difference between the two estimates; and the t statistics is calculated as shown above. The asterisks (*) show areas where the differences were statistically significant at different levels of significance. Three asterisks (***) in particular indicate that statistically the different is very important and cannot be ignored.

The assessment results show that most of the estimates are not statistically significantly different to the NATSISS values. However, there are some exceptions. The proportion of persons with a household member running out of money is different to the NATSISS estimates for some areas. The estimate for Tasmania/ACT (17.1%) and Victoria (27.6%) are significantly lower than NATSISS estimate (both at around six percentage point) while the estimate for the NT (32.8%) and Queensland (25.3%) are three and two percentage points higher, respectively, than the NATSISS estimate. Other variables that have a significant difference are participation in cultural activity and whether being a victim of a physical threat. The estimated value of cultural participation is significantly higher in the NT (79.8%) by four percentage points while significantly lower in WA (57.7%) by 7 percentage points. Tasmania/ACT is the only area that has a significant difference in the estimated proportion of victim of physical threat - the estimated proportion of 28% is three percentage points higher than the proportion estimated by NATSISS

Table 4 Comparison between the State estimates of the synthetic data and NATSISS

Name	Area	NATSISS (%)	Reweighted Synthetic (%)	Differences (% Point)	t-statistic
Participation in selected cultural activities	NSW	0.522	0.542	0.02	1.331
	Victoria	0.532	0.513	-0.02	-1.426
	Queensland	0.613	0.614	0	0.028
	SA	0.546	0.543	-0.003	-0.150
	WA	0.651	0.577	-0.074***	-5.252
	NT	0.758	0.798	0.04***	3.376
	Tasmania/ACT	0.606	0.621	0.014	0.906
Felt discriminated against	NSW	0.243	0.248	0.005	0.384
	Victoria	0.275	0.277	0.003	0.210
	Queensland	0.279	0.287	0.008	0.600
	SA	0.342	0.321	-0.021	-1.222
	WA	0.348	0.337	-0.011	-0.786
	NT	0.287	0.271	-0.016	-1.274
	Tasmania/ACT	0.160	0.164	0.005	0.403
Poor health status	NSW	0.095	0.092	-0.003	-0.287
	Victoria	0.081	0.082	0	0.026
	Queensland	0.052	0.055	0.003	0.465
	SA	0.082	0.073	-0.009	-0.914
	WA	0.061	0.061	0.001	0.071
	NT	0.074	0.074	0	-0.014
	Tasmania/ACT	0.074	0.074	0	0.024
High Kessler (K5) score	NSW	0.324	0.314	-0.01	-0.717
	Victoria	0.361	0.356	-0.005	-0.391
	Queensland	0.303	0.298	-0.005	-0.353
	SA	0.369	0.362	-0.007	-0.395
	WA	0.330	0.323	-0.007	-0.531
	NT	0.295	0.291	-0.004	-0.290
	Tasmania/ACT	0.297	0.289	-0.008	-0.504

Name	Area	NATSISS (%)	Reweighted Synthetic (%)	Differences (% Point)	t-statistic
Household members ran out of money for basic living expenses	NSW	0.314	0.319	0.005	0.383
	Victoria	0.335	0.276	-0.058***	-4.463
	Queensland	0.231	0.253	0.022*	1.791
	SA	0.325	0.326	0.002	0.094
	WA	0.321	0.301	-0.02	-1.474
	NT	0.296	0.328	0.032**	2.510
	Tasmania/ACT	0.230	0.171	-0.059***	-4.285
Whether victim of physical/threatened	NSW	0.241	0.248	0.007	0.581
	Victoria	0.284	0.282	-0.002	-0.192
	Queensland	0.204	0.213	0.01	0.814
	SA	0.271	0.256	-0.016	-0.968
	WA	0.301	0.287	-0.014	-1.059
	NT	0.235	0.231	-0.004	-0.363
	Tasmania/ACT	0.245	0.280	0.035**	2.507

*=significant at 10%, **=significant at 5%, *** = significant at 1%

4 Applying the spatial microsimulation model

The third step in the creation of the synthetic small area database is to reweight the national synthetic microdata using benchmark tables for small areas from the census. The benchmark tables are frequency tables or cross-tables derived from the Census and are derived for the SA2 geography. This process follows the spatial microsimulation method outlined in Tanton et al (2011).

4.1 Benchmark tables

Setting the benchmark tables is an important step in spatial microsimulation. The benchmark tables provide the information on particular characteristics of the population in small areas and assist in providing reliable estimates for the variable of interest. Therefore, the variables used for the benchmarks should be correlated with the variables of interest. Preliminary regression analysis as shown in section 3 often informs the choice of these variables (Anderson 2007; Chin and Harding 2006). Although Table 3 in Section 3 does inform the choice of variables in this study, the main deciding factor for choosing the benchmarks was to maximise the number of variables we could estimate with the final dataset. This is because in this study we are estimating several variables of interest using the spatial microsimulation process and it would not be efficient for us to produce different sets of weights for different variables estimated.

There are nine Census benchmark tables prepared for the spatial microsimulation model and these are listed in Table 5. The Census benchmark tables are derived from the newly released 2011 Census data from ABS Census Tablebuilder. Family type, age, sex, education, labour force status and housing tenure type are the priority benchmark tables since these variables appear most often in Table 3. Several benchmark tables use cross tabulations of these variables based on two criteria. The first is consultation on how a combination of variables could be important in revealing characteristics of local Indigenous populations. The second is looking at existing spatial microsimulation models that have been developed to estimate various measures at a small area level (for example Taylor et al., 2004; Chin et al., 2005; Harding et al., 2009; Tanton et al., 2009).

Table 5 Benchmark Tables

Number	Tables
1	Age by Sex by Labour Force Status
2	Number of children by Sex of Adult
3	Education (School and Non-school Qualification)
4	Indigenous Language by Non-school qualification
5	Tenure and Landlord composition
6	Household equivalised income
7	Family composition
8	Vehicle ownership
9	Occupation

Another important aspect in the spatial microsimulation method is the spatial unit used for the analysis. This spatial unit determines the level of spatial aggregation produced by the microsimulation. This study uses the SA2 as the base spatial unit of analysis. As discussed in Section 2, the SA2 is part of the new ABS standard geography (the ASGS). The main reason for the use of SA2's is that with around 2,200 spatial units, it is the smallest spatial unit within the ASGS which can be efficiently modelled using SpatialMSM. SA1's are smaller and less populous in terms of the Indigenous population, and thus harder to model with the SpatialMSM model. Another reason for the use of SA2's is that the SA2 geography covers the whole of Australia, unlike some other spatial units such as postcodes or local government authority areas (Mcnamara et al., 2009).

4.2 Results and validation from the spatial microsimulation

The estimation process using spatial microsimulation may not always produce a reliable result. Therefore, several quality assurance statistics are used to ensure that the result is acceptable. The first quality assurance measure used in this model is the Total Absolute Error (TAE) of the estimate from all the benchmarks. This statistic gives a measure of the accuracy of the reweighting process. The approach was suggested by Williamson et al., 1998, and is supported by other studies such as Smith et al., 2009; and Voas and Williamson, 2000. The design of the SpatialMSM model means that it automatically eliminates any area that fails to achieve the threshold of acceptable error (Tanton et al 2010). Those areas that fail this test are usually areas where the population is quite different to the sample population – for instance, industrial estates or inner city areas.

In the SpatialMSM model, the TAE is a measure of convergence for the model. With a larger number of benchmark tables, the GREGWT procedure used will have greater trouble converging, so reducing the number of benchmark tables will get estimates for areas where no estimate was available previously. Ideally, the number of benchmark tables will trade off getting the maximum number of areas with getting reliable estimates.

Table 6 shows the number and proportion of rejected SA2's using 7, 8 and 9 benchmarks. As explained above, it shows that a higher number of benchmark tables makes it more difficult for the estimation process to produce areas with TAE estimates below the threshold. Table 6 also shows the proportion of the Indigenous population excluded. While the table indicates that there are around 12 per cent of the SA2s excluded, it also indicates that the model omits between 2.5 and 3.4 per cent of the Indigenous population with the exception of Melbourne and the Australian Capital Territory.

For this analysis, we opted to use the results from the 8 benchmark tables as the number of areas gained by reducing the number of benchmark tables further is not much (only 17 SA2s or 0.8 per cent) while adding a benchmark would cause another 82 SA2's to be dropped from our analysis.

Table 6 Number of SA2 dropped due to higher Total Absolute Error (TAE) than threshold

	7 Benchmarks			8 Benchmarks			9 Benchmarks		
	Rejected number of SA2	% of SA2s in area	% of Indigenous population in area	Rejected number of SA2	% of SA2s in area	% of Indigenous population in area	Rejected number of SA2	% of SA2s in area	% of Indigenous population in area
Sydney	24	8.6	2.4	25	9.0	2.5	37	13.3	3.9
NSW-BS	3	1.2	0.4	3	1.2	0.4	4	1.5	0.5
Melbourne	92	32.7	12.4	97	34.5	13.2	119	42.3	17.3
Victoria-BS	22	14.5	3.3	23	15.1	3.5	34	22.4	6.0
Brisbane	20	8.5	3.3	21	8.9	3.4	30	12.7	4.0
Queensland-BS	9	3.1	1.9	9	3.1	1.9	12	4.1	2.3
Adelaide	18	16.5	4.0	21	19.3	8.0	25	22.9	9.1
SA-BS	10	16.4	2.6	10	16.4	2.6	17	27.9	4.9
Perth	21	12.1	5.9	24	13.9	6.3	26	15.0	6.7
WA-BS	4	5.2	0.8	4	5.2	0.8	5	6.5	0.9
Tasmania	1	1.0	0.2	2	2.0	0.3	4	4.1	0.8
NT	4	5.9	2.9	4	5.9	2.9	5	7.4	3.0
ACT	37	33.6	19.9	39	35.5	21.9	46	41.8	25.4
Australia	265	12.1	2.5	282	12.9	2.7	364	16.6	3.4

The other quality assurance measure is to validate the final estimates against a reliable source of small area data. This validation is a difficult task, as the whole purpose of producing these small area estimates is to create data where none currently exists. Two common approaches have been used to check the validity of these types of estimates. The first approach is to choose the closest possible variable to the estimated variable from a data source that is reliable for the chosen level of spatial disaggregation – for example, calculating poverty rates from the Census for each small area using an income cut-off available on the Census rather than the half median income cut-off traditionally used. Unfortunately, this option is not available for these estimates as there is no publically available small area data on the types of Indigenous disadvantage that we are estimating.

The second approach is to compare the estimate from the spatial microsimulation model to estimates for the smallest spatial area available reliably from the original survey. The smallest geographic area available from the NATSISS can be extracted using the Remote Access Data Laboratory (RADL) from the ABS. Using this application, reliable data for 17 areas based on State and Remoteness Area can be extracted.

This creates another issue as the SA2 geography that we are using for the spatial microsimulation does not exactly match these areas as parts of an SA2 can be located in different remoteness areas. Fortunately, most of the SA2's that are split across remoteness areas usually have a large part of their area categorised into one remoteness area and therefore, we can allocate the SA2 into the remoteness area that contains the greatest part of that SA2.

One issue that arises from this validation is that the estimation fails to capture fully the difference between remote and regional areas that has been captured in the NATSISS. For example, for the variable cultural participation of adults in Balance of State – Queensland, the model is able to estimate the different participation rates in different remoteness areas, but with a much smaller range than that shown in the NATSISS. The model estimates the cultural participation rate as between 56 per cent in inner regional areas to 67 per cent in remote areas, while the NATSISS indicates that the range of participation varies from 45 per cent in inner regional areas to 82 per cent in remote areas. This is because we only separated each state into capital city and balance of state, and could not separate the remote areas from the Balance of State.

The spatial microsimulation process requires weights that are reasonably accurate, and because we haven't been able to include remote areas in our calculation so far because they have been part of the Balance of State, we need to get better starting weights for remote areas. To do this, we benchmarked the synthetic data which was only for Capital City/Balance of State to aggregate survey data which included remoteness area.

Table 7 shows how the results from the spatial microsimulation with 8 benchmark tables compares with the results from the NATSISS for these larger geographies. In the first instance, the results seem to be reasonable – the results from the NATSISS for the 17 areas is close to the results from the aggregated spatial microsimulation process. The indicators provide different results – some are estimated very well (participation in cultural activities, feelings of discrimination, high Kessler and whether victim of physical/threatened violence), and some indicators are not estimated very well in some areas (poor health status in Sydney and Melbourne, ran out of money for basic living expenses in Melbourne, SA Non Remote, NT Remote and Very Remote).

The final stage of validation is an important one, and it is 'ground truthing' the results with a group of people who know the spatial patterns of Indigenous disadvantage. This includes experts on Indigenous disadvantage, and Indigenous groups themselves. This process is ongoing.

Table 7 Comparison between the spatial microsimulation estimates and NATSISS

Indicator	Area	NATSISS 2008	SpatialMSM 2011	Differences	t-statistics
Participation in selected cultural activities	NSW Major Cities	0.451	0.473	0.022	0.907
	NSW Inner Regional	0.551	0.578	0.027	0.979
	NSW Outer Regional	0.660	0.676	0.016	0.461
	Victoria Major Cities	0.477	0.499	0.022	1.189
	Victoria Inner/Outer Regional	0.593	0.627	0.034	1.204
	Queensland Major Cities	0.542	0.537	-0.004	-0.093
	Queensland Inner Regional	0.457	0.436	-0.021	-0.654
	Queensland Outer Regional	0.627	0.630	0.003	0.099
	Queensland Remote/Very Remote	0.820	0.789	-0.031	-1.380
	SA Non-Remote	0.523	0.535	0.012	0.579
	WA Major Cities	0.538	0.549	0.012	0.393
	WA Inner/Outer Regional	0.575	0.565	-0.01	-0.337
	WA Remote/Very Remote	0.788	0.793	0.005	0.260
	Tasmania Non-Remote	0.608	0.610	0.002	0.094
	NT Remote/Very Remote	0.785	0.780	-0.005	-0.348
	Balance of Australia - Non-remote	0.634	0.592	-0.042**	-2.001
	Balance of Australia - Remote/Very Remote	0.510	0.533	0.023	0.864
Felt discriminated against	NSW Major Cities	0.207	0.237	0.031	1.538
	NSW Inner Regional	0.270	0.282	0.012	0.507
	NSW Outer Regional	0.293	0.294	0	0.006
	Victoria Major Cities	0.252	0.265	0.013	0.780
	Victoria Inner/Outer Regional	0.305	0.322	0.017	0.625
	Queensland Major Cities	0.332	0.345	0.013	0.273
	Queensland Inner Regional	0.239	0.220	-0.019	-0.701
	Queensland Outer Regional	0.294	0.308	0.014	0.587
	Queensland Remote/Very Remote	0.268	0.297	0.029	1.113
	SA Non-Remote	0.320	0.328	0.008	0.406
	WA Major Cities	0.373	0.403	0.03	1.044
	WA Inner/Outer Regional	0.375	0.387	0.012	0.397
	WA Remote/Very Remote	0.285	0.286	0.001	0.055
	Tasmania Non-Remote	0.097	0.100	0.003	0.245
	NT Remote/Very Remote	0.261	0.284	0.023	1.639
	Balance of Australia - Non-remote	0.323	0.350	0.027	1.347
	Balance of Australia - Remote/Very Remote	0.217	0.238	0.021	0.926

Poor health status	NSW Major Cities	0.109	0.149	0.04***	2.615
	NSW Inner Regional	0.068	0.080	0.013	0.912
	NSW Outer Regional	0.115	0.127	0.013	0.537
	Victoria Major Cities	0.082	0.126	0.044***	4.306
	Victoria Inner/Outer Regional	0.078	0.101	0.022	1.426
	Queensland Major Cities	0.050	0.058	0.008	0.372
	Queensland Inner Regional	0.031	0.033	0.003	0.230
	Queensland Outer Regional	0.072	0.077	0.005	0.388
	Queensland Remote/Very Remote	0.047	0.057	0.01	0.817
	SA Non-Remote	0.088	0.107	0.019	1.597
	WA Major Cities	0.074	0.088	0.014	0.891
	WA Inner/Outer Regional	0.033	0.029	-0.004	-0.348
	WA Remote/Very Remote	0.064	0.058	-0.005	-0.472
	Tasmania Non-Remote	0.076	0.096	0.02*	1.778
	NT Remote/Very Remote	0.068	0.079	0.011	1.350
	Balance of Australia - Non-remote	0.092	0.106	0.014	1.109
	Balance of Australia - Remote/Very Remote	0.066	0.108	0.043***	3.207
High Kessler (K5) score	NSW Major Cities	0.316	0.281	-0.035	-1.533
	NSW Inner Regional	0.336	0.340	0.004	0.147
	NSW Outer Regional	0.296	0.284	-0.012	-0.360
	Victoria Major Cities	0.343	0.319	-0.024	-1.353
	Victoria Inner/Outer Regional	0.368	0.377	0.009	0.338
	Queensland Major Cities	0.321	0.314	-0.007	-0.163
	Queensland Inner Regional	0.218	0.206	-0.013	-0.479
	Queensland Outer Regional	0.357	0.367	0.009	0.377
	Queensland Remote/Very Remote	0.270	0.300	0.03	1.149
	SA Non-Remote	0.337	0.323	-0.014	-0.701
	WA Major Cities	0.306	0.324	0.018	0.647
	WA Inner/Outer Regional	0.319	0.301	-0.018	-0.602
	WA Remote/Very Remote	0.339	0.333	-0.007	-0.319
	Tasmania Non-Remote	0.277	0.300	0.023	1.236
	NT Remote/Very Remote	0.301	0.315	0.013	0.910
	Balance of Australia - Non-remote	0.263	0.258	-0.005	-0.252
	Balance of Australia - Remote/Very Remote	0.301	0.262	-0.039	-1.596

Household members ran out of money for basic living expenses	NSW Major Cities	0.300	0.298	-0.002	-0.094
	NSW Inner Regional	0.354	0.375	0.021	0.804
	NSW Outer Regional	0.243	0.257	0.014	0.427
	Victoria Major Cities	0.310	0.361	0.051***	3.007
	Victoria Inner/Outer Regional	0.309	0.340	0.031	1.168
	Queensland Major Cities	0.292	0.378	0.086*	1.950
	Queensland Inner Regional	0.125	0.116	-0.009	-0.454
	Queensland Outer Regional	0.303	0.332	0.028	1.176
	Queensland Remote/Very Remote	0.198	0.171	-0.027	-1.152
	SA Non-Remote	0.306	0.230	-0.076***	-3.913
	WA Major Cities	0.282	0.306	0.024	0.898
	WA Inner/Outer Regional	0.247	0.252	0.005	0.190
	WA Remote/Very Remote	0.294	0.293	0	-0.005
	Tasmania Non-Remote	0.179	0.200	0.02	1.285
	NT Remote/Very Remote	0.348	0.394	0.046***	3.014
	Balance of Australia - Non-remote	0.254	0.280	0.026	1.383
	Balance of Australia - Remote/Very Remote	0.242	0.242	0	-0.004
Whether victim of physical/threatened violence	NSW Major Cities	0.241	0.274	0.033	1.586
	NSW Inner Regional	0.273	0.278	0.005	0.193
	NSW Outer Regional	0.251	0.231	-0.02	-0.614
	Victoria Major Cities	0.278	0.298	0.02	1.216
	Victoria Inner/Outer Regional	0.292	0.306	0.014	0.542
	Queensland Major Cities	0.239	0.242	0.003	0.072
	Queensland Inner Regional	0.159	0.154	-0.005	-0.233
	Queensland Outer Regional	0.240	0.243	0.003	0.140
	Queensland Remote/Very Remote	0.193	0.194	0	0.013
	SA Non-Remote	0.264	0.237	-0.027	-1.429
	WA Major Cities	0.310	0.340	0.029	1.058
	WA Inner/Outer Regional	0.242	0.220	-0.022	-0.809
	WA Remote/Very Remote	0.288	0.289	0.001	0.050
	Tasmania Non-Remote	0.259	0.265	0.006	0.309
	NT Remote/Very Remote	0.205	0.205	0.001	0.069
	Balance of Australia - Non-remote	0.332	0.330	-0.002	-0.093
	Balance of Australia - Remote/Very Remote	0.191	0.190	0	-0.014

5 Concluding Remarks

The use of spatial microsimulation to derive small area estimates for a range of economic and social variables in Australia is increasing. It has been used by Government and non-Government institutions to help allocate services as well as by researchers. With the increasing effort by the Australian Government to reduce the gap in disadvantage between Indigenous and non-Indigenous communities, there is an increasing demand for small area data for Indigenous communities. Unfortunately, spatial microsimulation requires a record unit file from a survey, and this is not available for the Indigenous Social Survey that the ABS conducts due to issues with confidentiality. This study overcomes this issue by building a synthetic unit record file designed to match aggregate tables from the NATSISS.

The spatial microsimulation method produces small area estimates by reweighting the survey unit record data based on small area benchmarks provided from Census data. Having no available unit record data for Indigenous people means the main effort for this project was to create a synthetic unit record data set. This was then used as an input into the spatial microsimulation model. This work starts by creating synthetic data using probabilities taken from the Census data. This first stage produces a set of synthetic data with around 10 per cent error compared to the original table from the survey. This error can then be reduced to around 3 per cent by conducting a reweighting procedure.

The next step involves regression techniques to impute the variables that we would like to estimate from the NATSISS onto our dataset. The results from the regression analysis showed that better results were obtained running regressions separately by State and Capital City/Balance of State. The overall model results showed statistical significance, despite a low pseudo-R2, and the aggregation and validation of the modelled variables showed good results when validated against the aggregated survey data.

The creation of the synthetic unit record data has then allowed us to produce small area estimates of Indigenous disadvantage using spatial microsimulation. The validation of the results at an aggregate level indicates that we could produce an acceptable estimate using this technique although an alignment process was needed to ensure we could best capture the different characteristics of different communities. The final estimates can then be used by Indigenous communities and the Government for planning and service delivery purposes.

6 References

Anderson B (2007) Creating small area income deprivation estimates for Wales: Spatial microsimulation modelling. Chimera Working Paper 2007-11, Colchester, University of Essex.

Atkinson AB, Rainwater L, Smeeding TM (1995) Income distribution in OECD Countries: Evidence from Luxembourg Income Study, Social Policy Studies Vol. 18. Paris: OECD

Australia Bureau of Statistics (ABS) (2006*) ABS Remote Access Data Laboratory (RADL), User Manual*, Cat. 1406.0.55.002

Australia Bureau of Statistics (ABS) (2008) *National Aboriginal and Torres Strait Islander Social Survey: Users' Guide*, Cat. 4720.0

Australia Bureau of Statistics (ABS) (2010) *National Aboriginal and Torres Strait Islander Social Survey, 2008*, Cat. no. 4714.0

Australia Bureau of Statistics (ABS) (2011a) *Census Special Enumeration Strategies, Information Paper 2011*, Cat. 2911.0.55.004

Australia Bureau of Statistics (ABS) (2011b) *Main Structure and Greater Capital City Statistical Areas , Volume 1*, Cat. 1270.0.55.001

Ballas D, Clarke G, Dorling D, Eyre H, Thomas B, Rossiter D, (2005) SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place* 11 13- 34

Bell, P. (2000) *GREGWT and TABLE macros - Users guide*, Unpublished, Australian Bureau of Statistics

Caldwell S B, Clarke G P, Keister L A, (1998) Modelling regional changes in US household income and wealth: A research agenda, *Environment and Planning C: Government and Policy* 16 707-722.

Cassells R, Harding A, Miranti R, Tanton R, McNamara J, (2010) Spatial Microsimulation: Preparation of Sample Survey and Census Data for SpatialMSM/08 and SpatialMSM/09, Technical Paper No 36, National Centre for Social and Economic Modelling, Canberra

Chin S F, Harding A, (2007) SpatialMSM - NATSEM's small area household model for Australia in A. Gupta and A. Harding, (eds), *Modelling our future: Population ageing health and aged care*. Elsevier, Oxford, pp. 563 - 566

Chin S F, Harding A, Lloyd R, McNamara J, Phillips B, Vu Q N, (2005) Spatial microsimulation using synthetic small area estimates of income, tax and social security benefits *Australasian Journal of Regional Studies* 11 303- 336

Elbers C, Lanjouw JO, Lanjouw P (2003) Micro-level estimation of poverty and inequality. *Econometrica* 71 355–364

Fabrizi E., Giusti C, Salvati N, Tzavidis N (2013) Mapping average equivalized income using robust small area methods, *Papers in Regional Science*, DOI: 10.1111/pirs.12015

Ghosh M, Rao, J N K, (1994) Small Area Estimation: An Appraisal *Statistical Science* 9(1) 55-76

Pfeffermann D, (2002) Small area estimation - new developments and directions *International Statistical Review* 70(1) 125-143.

Harding A, and Tanton R, (2011) Policy and People at the Small Area Level: Using Microsimulation to Create Synthetic Spatial Data, in Stimson, R. *Handbook in Spatially Integrated Social Science Research Methods,* Edward Elgar, Sydney

Harding A, Gupta A, (2007) Introduction and Overview, in Harding, A and Gupta, A. (eds), *Modelling Our Future: Population Ageing, Social Security and Taxation*, Chapter 1, International Symposia in Economic Theory and Econometrics, Volume 15, Elsevier B. V., Amsterdam, pp. 1-29

Harding A, Vu Q N, Tanton R and Vidyattama Y, (2009) Improving work incentives for parents: the national and geographic impact of liberalising the Family Tax Benefit income test *The Economic Record*, 85 (Special Issue) 48 – 58

Harding, A, Vidyattama, Y & Tanton, R, (2011) Demographic Change and the Needs-Based Planning of Government Services: Projecting Small Area Populations Using Spatial Microsimulation *The Journal of Population Research*, 28(2-3) 203-224

Hermes K, and Poulsen M, (2013) The intraurban geography of generalised trust in Sydney. *Environment and Planning A*, 45(2), 276-294.

Hynes S, Morrissey K, O'Donoghue C, Clarke G, (2009) Building a static farm level spatial microsimulation model for rural development and agricultural policy analysis in Ireland, *International Journal of Agricultural Resources, Governance and Ecology* 8(2) 282-299

Kennedy B, and Firman D, (2004) Indigenous SEIFA – revealing the ecological fallacy. Paper for the 12th Biennial Conference of the Australian Population Association, 15 - 17 September 2004, Canberra.

Lymer S, Brown L, Harding A, Yap M, (2009) Predicting the need for aged care services at the small area level: the CAREMOD spatial microsimulation model *International Journal of Microsimulation* 2(2), 27-42

Morphy F, Sanders W, and Taylor J, (2007) Accommodating agency and contingency: towards an extended strategy for engagement in Frances Morphy (Editor) *Agency, contingency and census process: Observations of the 2006 Indigenous Enumeration Strategy in remote Aboriginal Australia*, Research Monograph No. 28, Centre for Aboriginal Economic Policy Research The Australian National University, Canberra, pages 113-125

Nakaya T, Fotheringham A S, Hanaoka K, Clarke G P, Ballas D, Yano K, (2007) Combining microsimulation and spatial interaction models for retail location analysis *Journal of Geographical Systems* 9(4) 345-369

Pratesi M, Salvati N (2008) Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications* 17(1):113–141

Rahman A, Harding A, Tanton R, Liu S, (2010) Methodological Issues in Spatial Microsimulation Modelling for Small Area Estimation *International Journal of Microsimulation* 3(2), 3-22

Smith D M, Clarke G P, Harland K, 2009, Improving the synthetic data generation process in spatial microsimulation models *Environment and Planning A*, 41, 1251-1268

Tanton R, Vidyattama Y, McNamara J, Vu Q N, Harding A, (2009) 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* 28(2), 102 – 120

Tanton R, Vidyattama Y, Nepal B, McNamara J, (2011) Small area estimation using a reweighting algorithm *Journal of the Royal Statistical Society Series A (Statistics in Society)* 174(4) 931-951

Tanton R, Williamson P, Harding A, (2007) Comparing two methods of reweighting a survey file to small area data: generalised regression and combinatorial optimisation 1st General Conference of the International Microsimulation Association, Vienna, Austria

Tanton, R., & Edwards, K. L. (2013). *Spatial Microsimulation: A Reference Guide for Users*. Springer Netherlands.

Taylor E, Harding A, Lloyd R and Blake M. (2004) Housing Unaffordability at the Statistical Local Area Level: New Estimates Using Spatial Microsimulation, *Australasian Journal of Regional Studies*, 10(3), 279-300

van Leeuwen E, Clarke G P, Rietveld P, (2009) Microsimulation as a tool in spatial decision making: simulation of retail developments in a Dutch town in Zaidi A, Harding A and Williamson P (Eds.), *New frontiers in microsimulation modelling*, Ashgate, Aldershot, pp 97 - 122

Vidyattama Y, Cassells R, Harding A, McNamara J, (2011) Rich or Poor in Retirement? A Small Area Analysis of Australian Private Superannuation Savings in 2006 using Spatial Microsimulation *Regional Studies* 47(5), 722-739

Voas D and Williamson P, (2000) An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata *International Journal of Population Geography*, 6, 349 – 366

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.

Williamson P, (2013) An Evaluation of Two Synthetic Small-Area Microdata Simulation Methodologies: Synthetic Reconstruction and Combinatorial Optimisation in R Tanton and K Edwards (Eds) *Spatial Microsimulation: A Reference Guide for Users*, pp 19-47

Williamson P, Birkin M and Rees P, (1998) The estimation of population microdata by using data from small area statistics and samples of anonymised records *Environment and Planning A,* 30(5), 785-816

Zaidi A, Harding A, Williamson P, 2009, (eds), *New Frontiers in Microsimulation Modelling*, Ashgate, Aldershot.