AN EVALUATION OF SYNTHETIC HOUSEHOLD POPULATIONS FOR CENSUS COLLECTION DISTRICTS CREATED USING OPTIMISATION TECHNIQUES

Tony Melhuish
National Centre for Social and Economic Modelling (NATSEM), University of Canberra, Canberra, ACT 2601, Australia.

Marcus Blake
National Centre for Social and Economic Modelling (NATSEM), University of Canberra, Canberra, ACT 2601, Australia.

Susan Day
National Centre for Social and Economic Modelling (NATSEM), University of Canberra, Canberra, ACT 2601, Australia.

ABSTRACT: Regional policy makers rely on the availability of detailed and current small area data to inform their decision making. The main source of small area socio-demographic data in Australia is the five yearly Census of Population and Housing conducted by the Australian Bureau of Statistics (ABS). Although the Census provides a comprehensive coverage of Australian households, only a relatively limited range of information is collected about these households. Other data sources, such as the ABS Household Expenditure Survey (HES), provide a rich source of household information but are not available for small geographic areas. One solution to this lack of detailed small area data is to combine the information-rich survey data with the geographically disaggregated Census data using synthetic estimation. This paper reports on some initial validation of recent research that uses optimisation techniques to create a set of synthetic households that seek to represent the socio-demographic profiles of each Census Collection District (CCD) in Australia. The ABS Household Expenditure Survey Confidentialised Unit Record File is reweighted for each CCD to match selected variables in the Census Basic Community Profile (BCP). As an initial measure of the success of this technique, the weights generated are used to recreate the socio-demographic profiles of CCDs in the Australian Capital Territory. The socio-demographic profiles created by synthetic household populations are compared with those from the Census BCP to evaluate the degree to which the synthetic households represent the population within each CCD. This research demonstrates the potential for using optimisation techniques to create synthetic data that can be used for small area policy and household level analysis.

1. INTRODUCTION

The main purpose of this paper is to report on initial validation of a method for creating synthetic small area socio-demographic data, using a technique referred to as ‘synthetic estimation’. The first section of the paper describes the main sources of socio-demographic data currently available and the limitations of this data. The second section describes synthetic estimation and introduces the major methods of creating synthetic microdata. The synthetic estimation approach currently being developed by NATSEM, known as SYNAGI - Synthetic Australian Geo-demographic Information – is then described. Section 3 reports on the initial evaluation of the process used by the SYNAGI method by comparing the simulated socio-demographic characteristics of small areas with the same characteristics from the 1996 Census. The fourth section discusses the key assumption of this type of modelling and other important issues that need to be considered in interpreting the model results. The paper concludes with an assessment of the validation and suggestions for future research.

1.1 Existing Individual and Household Data

Regional policy makers rely on the availability of detailed and current small area data to inform their decision making. The main source of small area socio-demographic data in Australia is the five yearly Census of Population and Housing conducted by the Australian Bureau of Statistics (ABS). The Census is a count of the population and dwellings in Australia with details of age, sex and a variety of other characteristics (ABS, 1996). The smallest geographic area defined in the Census is the Census Collection District (CCD) which is used for collection, processing and output of data. There are approximately 225 dwellings in each urban CCD, with fewer dwellings in rural areas. There were a total of 34,410 CCDs defined in the 1996 Census[1].

In addition to the Census, the ABS conducts surveys to collect detailed information on incomes, expenditures and other individual and household characteristics, such as the Household Expenditure Survey (HES), the Survey of Income and Housing Costs (SIHC) and the National Health Survey (NHS). Household and individual information is also collected by numerous public and private agencies in the conduct of their day-to-day activities. These administrative data can contain vast amounts of information on an individual’s spending patterns, health history, travel habits and many other preferences, choices and characteristics. The results of market and attitudinal surveys are also a rich source of information that have the potential to contribute to corporate and public decision making.

1.2 Microdata

Microdata are data that are available at the unit record level and generally consist of a list of unidentifiable individuals or households with associated characteristics obtained from a survey or Census. Individual and household characteristics may include age, sex, marital status, household type, dwelling type and possibly a spatial indicator identifying the broad geographic location of the individual or household.

Microdata are available from the ABS from the Census and many of its surveys in the form of Confidentialised Unit Record Files (CURFs). Census microdata are available as a 1% Household Sample File of the Census population, with some levels of detail collapsed for confidentiality. CURFs are also available from the HES and SIHC, again with the measures taken by the ABS to ensure confidentiality. These CURFs contain all unit records included in each survey. CURFs provide a valuable source of unit record data and provide an

[1] In the 2001 Census the number of CCDs increased to 37,209.

opportunity for analysis at the individual or household level not available from tabular output. Usage of all CURFs is strictly governed by a licensing agreement with the ABS.

1.3 Limitations of Existing Data

Although the Census provides a comprehensive coverage of Australian households for small geographic areas, it has several major limitations. These include:

- the amount of information collected from each household is relatively limited. For example, only gross household income is collected and then only in broad ranges of income, and there is no information about social security receipt, income sources, wealth and expenditure;
- unlike many other ABS collections, the full Census results are not publicly available as a unit record file. Output for the whole Census file is only available as a pre-defined series of tables for each CCD, or as customised tables that can be purchased from the ABS. This means, for example, that relationships between characteristics of interest cannot be easily or fully explored (such as age by income by educational qualifications). It also means that traditional microsimulation models[2] – that are widely used by policy makers to assess the likely impact of policy changes on certain groups in society – cannot be constructed on top of the pre-defined tables; and
- to protect the confidentiality of individuals, the ABS randomises small numbers within the Census. This makes analysis of multiple characteristics for individuals or households unreliable for many small geographic areas.

Other ABS data sources, such as the Household Expenditure Survey, provide a rich source of household information but are not available for small geographic areas. Due to relatively small sample sizes and the limited spatial stratification of these surveys, very little information is available about the spatial variation of individual or household characteristics.

The major limitations of administrative and market survey data include their limited availability, difficulty in use (most data are not collected for analytical purposes and therefore can be difficult to process, particularly geographically) and reliability.

1.4. Synthetic Microdata

One solution to this lack of detailed small area data is to merge the information-rich survey data with the geographically disaggregated Census data to create synthetic microdata for small areas. This new data may then help to fill the deficiency in the information available to policy makers by providing synthetic small area unit record data – effectively by creating 225 or so synthetic households for each CCD whose characteristics match as closely as possible the

[2] Microsimulation models traditionally use microdata to estimate the likely overall impact of social or economic policy change on individuals or households by applying a set of rules to the individuals in the microdata. They are particularly useful for the analysis of the distribution of outcomes within the population rather than just aggregate outcomes.

characteristics of the 225 households living in that CCD as shown in the Census data[3] .

The benefits of creating synthetic microdata include:

- the creation of spatially disaggregated data from aggregated data such as national surveys;
- the ability to create tables of Census variables that are not available in the Census Basic Community Profiles (BCPs);
- using the many simulated characteristics of each individual or household for multivariate analysis, thereby providing a method of identifying and analysing specific socio-demographic groups at the small area level; and
- the potential to use traditional microsimulation models to estimate the spatial impact of policy on particular groups within the population.

2. METHODOLOGY

2.1 Synthetic Estimation

Estimates of the population can be direct or synthetic. A direct estimate uses data collected exclusively from the area for which the population is being estimated. Synthetic estimators generally use data from a larger region to make estimates about smaller areas within that region. In this research the term synthetic estimation is used to describe those techniques that create synthetic microdata for small geographic areas. These techniques generally rely on creating synthetic individuals or households that match the socio-demographic characteristics of the small areas of interest; the two major methods used to achieve this are discussed below. These techniques are spatial in that they rely on the geographic variation of the socio-demographic characteristics used in the creation of the synthetic data. Other spatial issues, particularly location specific influences such as accessibility or other ‘spatial’ effects (see section 4.2 below), are less explicit in this type of modelling methodology and are generally considered separately from the creation of the synthetic microdata.

2.2 Creating Microdata

Synthetic estimation is a technique that combines individual or household microdata, currently available only for large spatial areas, with small area data to create synthetic microdata estimates for these small areas. There are two possible methods by which this can be achieved - ‘synthetic reconstruction’ or ‘reweighting’ (Williamson et al, 1998).

The synthetic reconstruction approach requires the creation of a set of synthetic individuals or households whose characteristics match aggregate characteristics for the small area, such as those in the Census BCP tables. The process usually involves imputing characteristics based on the distributions within the constraining tables, building the individual or household profile in a sequential manner.

Reweighting is achieved by altering the weights for each individual or

[3] It should be noted that to allay any privacy concerns NATSEM does not allow external access to the individual simulated household records.

household in the survey. As national sample surveys are based on a sample of the population, each individual or household within the survey must be weighted to represent the total number of that type of household within the population (sometimes also called ‘grossing up’). In a similar manner, the same sample can be reweighted so it represents the population within a small area. This can be achieved by selecting a representative set of individuals or households that, when viewed together, best fit the aggregate characteristics of the small area. One way of doing so is to select 225 or so households from the sample survey that best represent a particular CCD (this is an integer method of selection, in which all selected households have a weight of one). Alternatively, all households within the sample can be given a small fractional weight so that the sum of all weights equals the population in the selected CCD and the sum of the fractional individuals or households best matches the characteristic profile of the CCD.

2.3 The SYNAGI Reweighting Approach

The SYNAGI (SYNthetic Australian Geo-demographic Information) approach currently being developed by NATSEM uses the reweighting method to blend the Census and ABS sample survey data together to create a synthetic unit record file for every CCD in Australia. To date, NATSEM’s efforts have focussed upon the ABS Household Expenditure Survey, although efforts are currently underway to extend the methodology to enable the ‘regionalisation’ of other sample survey data. The existing model first recodes the HES and Census variables to be comparable, and then reweights the HES, utilising detailed socio-demographic profiles from the Census BCPs. Reweighting is undertaken using an optimisation approach to iteratively generate a set of weights that ‘best-fits’ each CCD. That is, household weights are gradually changed until they produce a set of characteristics that match those of each CCD. Although a non-integer method of reweighting is used, in effect, the modelling can be seen as creating 225 or so synthetic households for each CCD, with the characteristics of the synthetic households within each CCD closely matching the characteristics revealed in the Census data for households in that particular CCD.

SYNAGI reweighting currently uses data from the 1996 Census of Population and Housing BCP to create target variables for each of the 34,410 CCDs in Australia. The variables from the Census that are chosen as targets are those that are also contained within the 1998-99 HES. To make the variables from the HES compatible with the Census, relevant HES variables are recoded so that they match the classifications and ranges that exist in the Census. A total of 15 variables are currently used in the SYNAGI matching process. These variables are listed in Table 1.

Within these 15 variables, 64 targets are defined by ranges within each of the broader characteristic groups; for example, 7 income ranges (X1 – X7), 10 age ranges (X8 – X17) and 3 broad regions of birth (X21 – X23). Most of these target variables are single variables but some are multivariate (such as high income segments by age (X41 - X43)).

Table 1. Variables Used in SYNAGI Reweighting.

Characteristic Group	Targets
Total household income	X1 – X7
Age	X8 – X17
Marital status	X18 – X20
Country of birth	X21 – X23
Labour force status by sex	X24 – X31
Occupation	X32 – X36
Family type	X37 - X38
Student status	X39 – X40
High income segments by age	X41 – X43
Housing type	X44 – X46
Housing tenure	X47 – X50
Household size	X51 - X55
Number of motor vehicles	X56 – X58
Mortgage repayments	X59 – X61
Rent payments	X62 – X64

The matching process requires that the Census and HES variables are based on the same year. This requires that the target variables from the 1996 Census are updated to the year of interest. Monetary values must be inflated and the population adjusted for each CCD, currently by using ABS building approvals data[4]. Similarly, HES data are also inflated. There is no requirement to increase the population size of the HES as it is a sample and is reweighted in the SYNAGI process to match the population within each CCD.

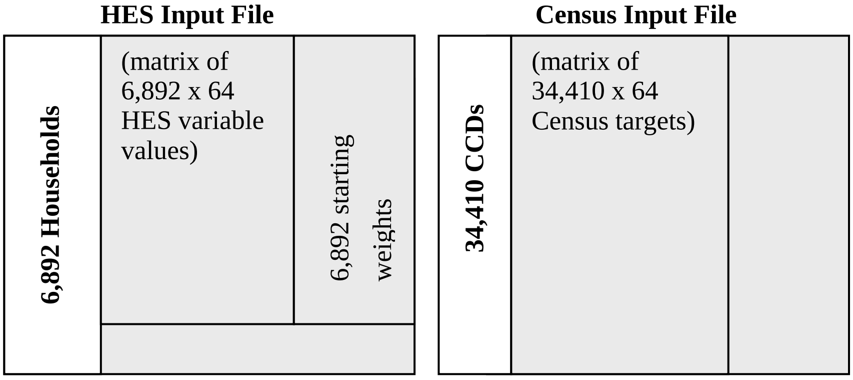
The current version of the SYNAGI model is based on the 1996 Census and the 1998-99 HES, with both data sources updated to June 2000. With the release of the 2001 Census data, the model will be updated to 2001 using the 2001 Census data.

After recoding to create consistent classifications and ranges the two datasets used in the reweighting process have the structure shown in Table 2.

The objective of the optimisation process is to reweight the HES households in an iterative manner to create a match for the target variables in the Census for each CCD. This results in a set of 6,892 household weights for each of the 34,410 CCDs, although many of the weights within a particular CCD will be zero. The sum of these weights equals the number of households in the CCD, while applying the weights to the 64 HES input values should create values that match the target values in the Census table.

[4] The method of updating Census variables in the current approach is fairly crude. As SYNAGI develops, methods will be developed to improve the estimation of population change for small areas and to estimate the likely change in the characteristics of these small areas. Given the complexities of change at the small area level, even between Censuses, this task is far from trivial and would rely on ancillary data, such as labour force estimates, to inform the updating process.

Table 2. Input Datasets used in SYNAGI Reweighting.



HES Input File

6,892 Households

(matrix of 6,892 x 64 HES variable values)

6,892 starting weights

Census Input File

34,410 CCDs

(matrix of 34,410 x 64 Census targets)

The optimisation process is a Fortran program consisting of three linked convergence algorithms that marginally change the values of household weights and subsequently evaluate the change in the 64 variable values compared with the Census targets. The first of the three algorithms focuses on each target sequentially with the aim of moving the weights and the resultant synthetic population closer to all target variables. These weights are then used in the second algorithm that has a more global focus, evaluating the overall fit to all targets. The final algorithm involves a multi-dimensional search for convergence by changing a pair of household weights in a positive and/or negative direction.

The evaluation measure is the absolute residual between each of the 64 reweighted HES values and the Census targets. In general terms, if the change in household weights improves the fit to the Census targets the weights are accepted, otherwise the change in weights is rejected. This process is undertaken many times until the reweighted HES values *converge* on the Census targets.

3. CREATING SYNTHETIC HOUSEHOLDS FOR THE AUSTRALIAN CAPITAL TERRITORY

For the validation exercise, weights have been generated for a subset of CCDs; namely, those in the ACT. Although the ACT is not representative of the whole of Australia in many of its characteristics, it is a manageable set of 492 CCDs that provides an opportunity to validate the method used in generating household weights.

3.1 Evaluation of Synthetic Microdata

Given that one of the objectives of creating synthetic microdata is to create data that does not currently exist for small geographic areas, validation of the results is difficult. For this reason, validation of the SYNAGI results relies heavily on the internal consistency of the model process. Similar to Voas and Williamson (2000), the model outcomes are assessed as a first stage validation in terms of their overall match to the socio-demographic profiles within each CCD. The performance of individual variables and the 64 targets is then assessed. A preliminary evaluation is also undertaken of the creation of a ‘new’ cross-tabulation between two of the simulated variables.

To reiterate, the reweighting algorithm central to the SYNAGI model generates a set of household weights that, when applied to the HES, seek to represent the socio-demographic profiles of individual CCDs. The algorithm in the model uses a vector of 64 Census derived targets against which the weights are 'optimised'. The starting weights for each CCD (ABS HES weights scaled down to total the number of households in the CCD) are iteratively adjusted until 64 simulated target values, produced by applying the weights to the HES, converge to the values of the 64 Census derived targets. As a basic validation of the optimisation process the success of this convergence is assessed.

The ACT weights used in the validation are based on the 1996 Census and an earlier version of the HES, updated to 1996. This was done to avoid introducing errors associated with updating the Census targets and to enable the validation of the optimisation process in isolation of the updating procedure.

Given that the resultant household weights for each CCD are meant to represent the households within that CCD, the simulated socio-demographic profile generated from the optimal set of weights should closely resemble the profile derived from the Census. In addition, the value of each of the 64 individual simulated values should closely match its equivalent Census derived target.

Achieving a good result for matching individual targets or creating a reasonable match to overall socio-demographic profiles would provide a level of confidence that the weights can be used to create 'new' data such as multivariate data not included in the BCPs. This household level multivariate data could then be used to ascribe other attributes to households not included in the optimisation process. Validation of this new and ascribed data needs to be addressed separately to determine the applications and issues that may benefit most from the current SYNAGI approach, as discussed later in this paper.

3.2 Socio-demographic Profiles

As a general measure as to whether the simulated CCD profiles match the Census derived profiles, the values for all 64 Census derived and simulated target values can be compared. Figure 1 shows the socio-demographic profile of one CCD in the ACT and illustrates the general magnitudes and comparability of the Census target values and the SYNAGI simulated values. In general, Figure 1 is representative of most CCDs in the ACT, as most CCDs display a very high level of comparability between Census and simulated profiles. Although there are a few exceptions, the overall profiles indicate that the values of simulated and Census variables converge extremely well and that all of the targets are well matched, both in terms of their overall distribution and the magnitude of their values.

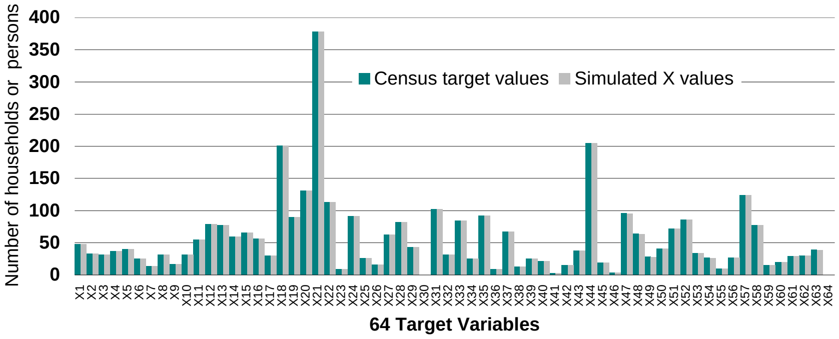


Figure 1. Census Target and Simulated Socio-demographic Profile for one CCD in the ACT.

Data Source: NATSEM & ABS 1996 Census data.

Note: See Table 1 for a description of the target variables.

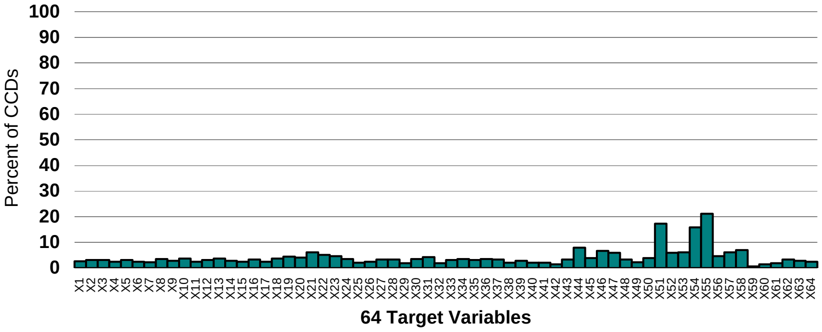


Figure 2. Proportion of CCDs in the ACT with Absolute Residuals Greater than One Household or Person.

Data Source: NATSEM & ABS 1996 Census data.
Note: See Table 1 for a description of the target variables.

3.3. Individual Target Variables

As an indicator of how well individual variables are being matched by the optimisation process, the absolute value of the difference between the simulated and Census values for each variable (the absolute residual) has been determined. The proportion of CCDs in the ACT in which the absolute residual for each of the 64 target variables is greater than one (that is, one household or person) is shown in Figure 2.

The majority of the targets appear to have converged very effectively for the majority of CCDs. In fact, 61 of the 64 simulated values converge to within one household or person of their respective Census targets in over 90% of CCDs.

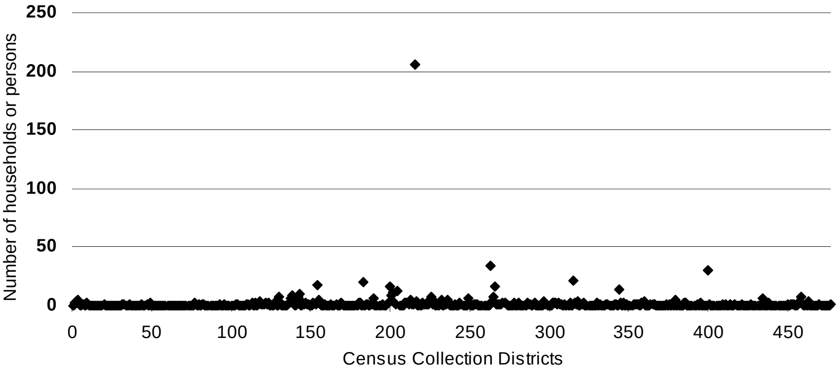


Figure 3. Maximum Residuals for Each CCD in the ACT.

Data Source*:* NATSEM & ABS 1996 Census data.

The only targets displaying slightly poorer convergence - that is, residuals greater than one - are household size targets X51 (lone person households), X54 and X55 (four and five or more person households, respectively). However, most CCDs converge within two households or persons for all targets, including household size.

3.4. Maximum Residuals

The absolute residuals are used by the reweighting algorithm to evaluate the success of the optimisation process. Of particular interest are the maximum residual and the sum of absolute residuals for a CCD. The maximum residual identifies the variable of least convergence and the sum of absolute residuals provides an overall measure of convergence.

Figure 3 is a scatter plot of maximum residuals for the CCDs in the ACT. It is apparent that most CCDs have a very low maximum residual (households or persons depending on the variable) with the majority less than 0.5. This is not surprising given that the optimisation process converges well on most variables and one of the convergence criteria is to achieve residuals of less than 0.5 for each target. If all targets for all CCDs in the ACT are considered, over 80% of the simulated values are within 1% of their target value - an extremely good outcome.

A small group of CCDs has a maximum residual greater than one with one greater than 200 (the ‘manager’ target X32 within the occupation variable). This large residual is in a CCD that has a generally poor fit due to its unusual socio-demographic profile.

As Figure 4 illustrates, those CCDs with a relatively large maximum residual compared with the convergence criterion of 0.5 are generally those that are unlikely to be exclusively urban-residential in character. Examples include: CCDs on the urban fringe where CCD profiles may contain urban and non-urban households, land use may include rural or semi-rural activities, and the number of households may be fewer than in established urban areas; and CCDs in the

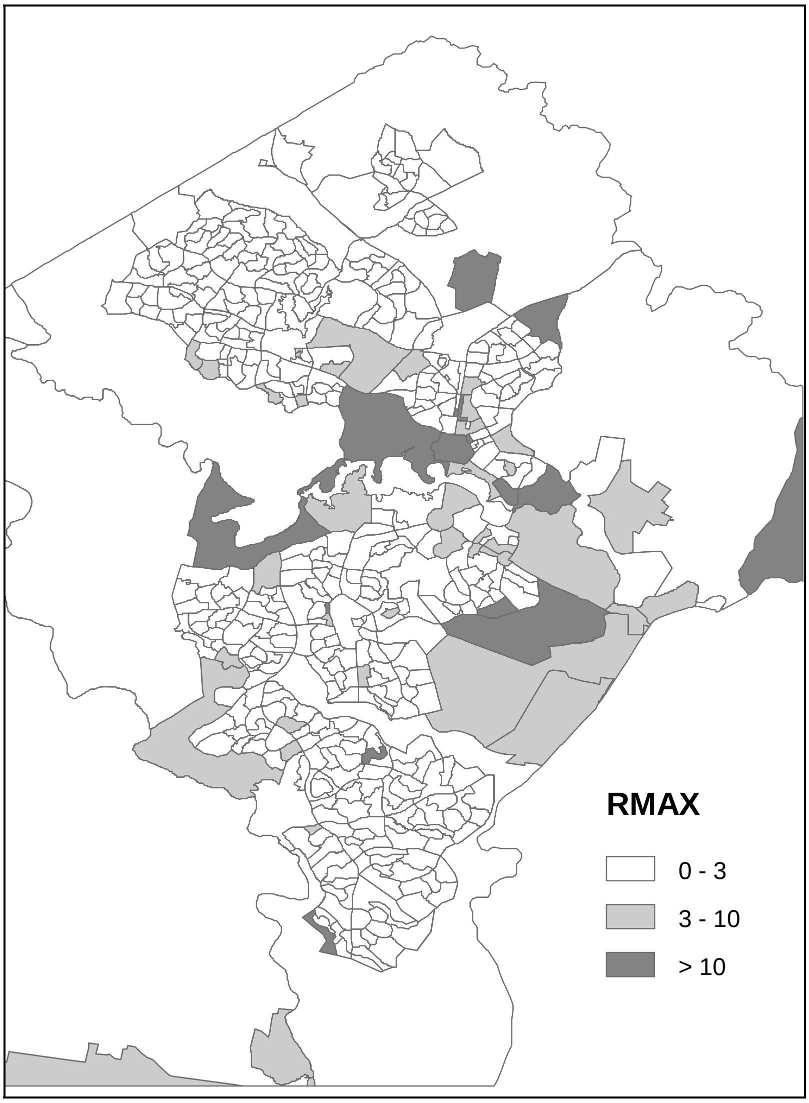


Figure 4. CCDs in the ACT with Maximum Residuals Greater than Three Households or Persons.

Data Source*:* NATSEM & ABS 1996 Census data.

city centre that have a commercial-residential mix with many people living in apartment style accommodation. These potentially unusual household characteristics and lower household numbers make convergence difficult – a finding also reported by Voas and Williamson (2000).

In addition to the urban fringe, other CCDs in the ACT that have high maximum residuals include the Australian National University north of Lake Burley Griffin, Civic Centre (the CBD of Canberra) and the Royal Military

College at Duntroon, east of the lake. Other than these ‘unusual’ CCDs, the majority of CCDs converge well on all variables.

3.5. Non-convergent CCDs

As a measure of non-convergence, it is interesting to consider those CCDs for which the overall evaluation of convergence is poor - that is, the sum of absolute residuals is the greatest. Figure 5 shows the simulated and Census values for the CCD with the greatest sum of absolute residuals. The character of this CCD is not typical of most CCDs in the ACT because of the nature of its inhabitants - defence personnel at the Duntroon military college. Clearly, many of the simulated values are substantially different from their target values resulting in a very poor fit overall. Of particular interest are the income (X1 – X7), occupation (X32 – X 36), student status (X39 – X40), housing type (X44 - X46), housing tenure (47 –X50) and mortgage repayment (X59 – X61) variables. The Census target values indicate that this is a CCD in which most households have a weekly income between $700 and $1,499, identify themselves as managers, are predominantly full-time students and live in detached, privately rented houses. This profile initially appears incongruous as, generally, full-time students would be expected to have low incomes and are unlikely to have managerial occupations. Equally, if a household has a reasonably high income, they may also be expected to be buying their own home. However, defence personnel residing at a military college are likely to be undertaking full-time study, have reasonable incomes and be living in defence housing. As SYNAGI is based on the HES, a sample survey, it is unlikely that many, if any, households selected for the survey have characteristics matching those of Duntroon, particularly as the HES only includes occupied *private* dwellings (that is, institutional dwellings are excluded). Therefore, the optimisation process is unlikely to find a set of weights that can match all 64 targets successfully.

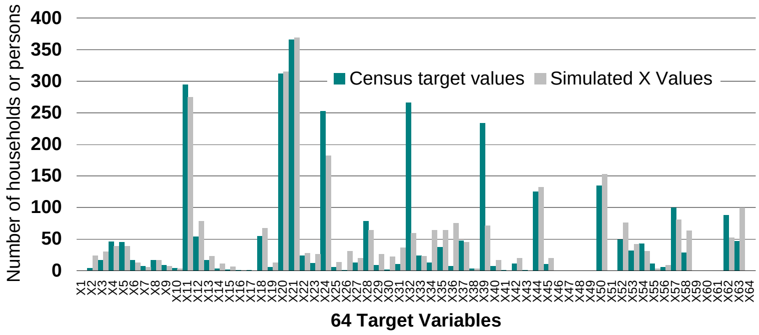


Figure 5. Census Targets and Simulated Socio-demographic Profiles for the Least Convergent CCD in the ACT.

Data Source*:* NATSEM & ABS 1996 Census data.
Note: See Table 1 for a description of the target variables.

Table 3. Percent Difference Between the Simulated and Census Totals for Tenure Type by Dwelling Structure for the ACT.

	Fully Owned	**Being Purchased**	**Rented**	**Total**
Separate house	2.3%	3.4%	6.2%	3.7%
Semi-detached, row or terraced house	5.3%	6.4%	4.4%	5.0%
Flat, unit or apartment	30.7%	10.6%	0.8%	4.4%
Total	3.3%	3.7%	4.6%	3.9%

Source: Derived from NATSEM & ABS 1996 Census data.

Generally, non-convergent CCDs contain large maximum residuals for individual Census targets and their distribution is very similar to those CCDs identified in Figure 4. That is, they are CCDs that are on the urban fringe, institutional CCDs such as the Australian National University, or the CBD.

3.6 Cross-tabulated Validation

As previously mentioned, most of the 64 Census targets are single variables, with the exception of several that are broken down by age or sex. In addition, the optimisation process is designed to produce a set of optimal weights that, when applied to HES values best match the socio-demographic profile of the 64 targets for each CCD. Assuming that the process is successful in matching all 64 targets for a CCD, how well does the set of household weights replicate the ‘true’ joint characteristics within the CCD? That is, can sub-groups within the CCD, such as low-income families living in rental accommodation, be identified successfully using the reweighting method? As you will recall, one of the benefits of creating simulated microdata is to enable the creation of multivariate data for small areas that does not currently exist - for example, Census cross-tabulations that are not currently published by the ABS. For this reason, there are few data sets available against which the simulated cross-tabulated microdata can be validated.

There is one BCP table in the 1996 Census that has not been used in the reweighting process - dwelling type by tenure type. This cross-tabulation has been recreated using the SYNAGI weights for the ACT. Table 3 compares the simulated results with the Census results for a 3 by 3 cross-tabulation of housing tenure and dwelling type for the whole of the ACT. As can be seen when the results for all 492 CCDs in the ACT are summed, the modelling produces a good fit for the majority of the table, but these preliminary results suggest that the tenure of flats is particularly hard to simulate.

Figure 6 illustrates the spatial distribution by CCD of differences between the simulated number of fully owned flats and the number in the 1996 ABS Census. Not surprisingly, those CCDs with the largest errors are generally those with the largest number of flats, mainly in the inner urban areas. Although this would suggest that proportional errors would be more relevant, the SYNAGI model currently uses the actual number of households or person in the convergence criteria.

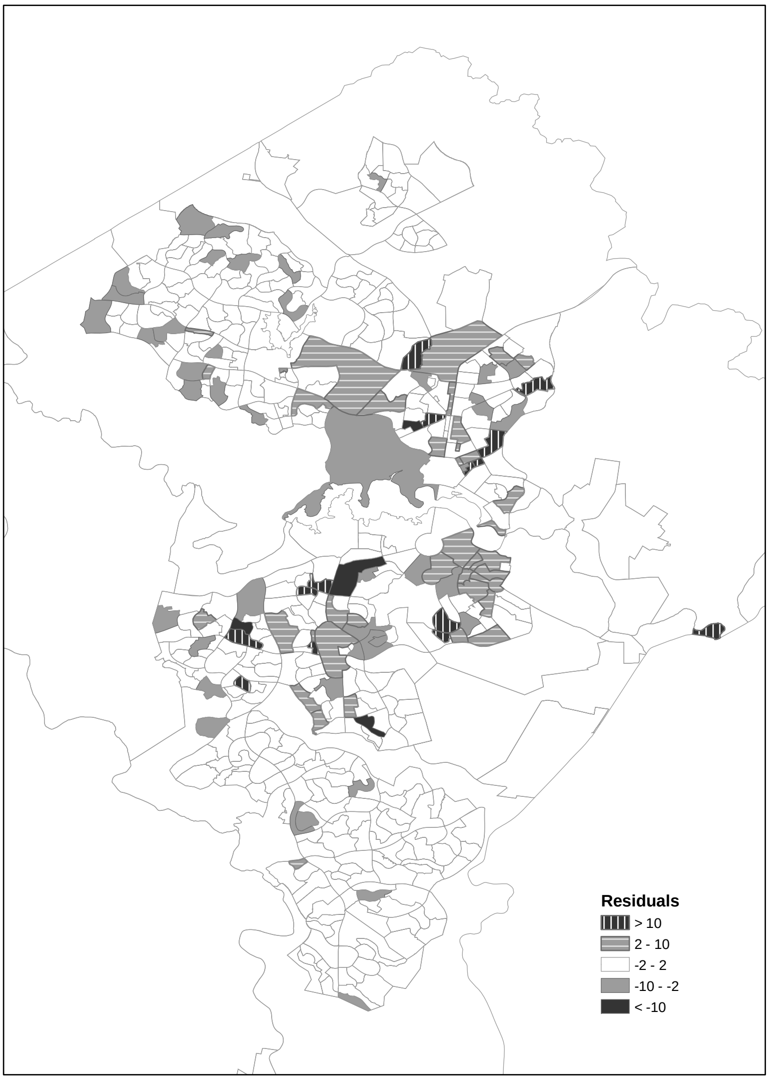


Figure 6. Difference Between the Number of Simulated and Actual Fully-Owned Flats for CCDs in the ACT.

Data Source: NATSEM.

To identify those CCDs that do not result in good cross-tabulated results, each CCD was scored according to how well each of the 9 table cells matched the Census targets. Criteria for both absolute and relative differences were set, and each CCD was given a score of one if the difference was greater than 5 households AND greater than 20 percent of the Census value, otherwise the CCD was given a score of zero. These cell scores were summed across the table giving a possible minimum score of zero and maximum score of nine. Figure 3.7 illustrates the spatial distribution of this scoring system. There appears to be little clustering of the results, with the CCDs with the highest scores generally dispersed throughout the urban areas of Canberra. Future work will include

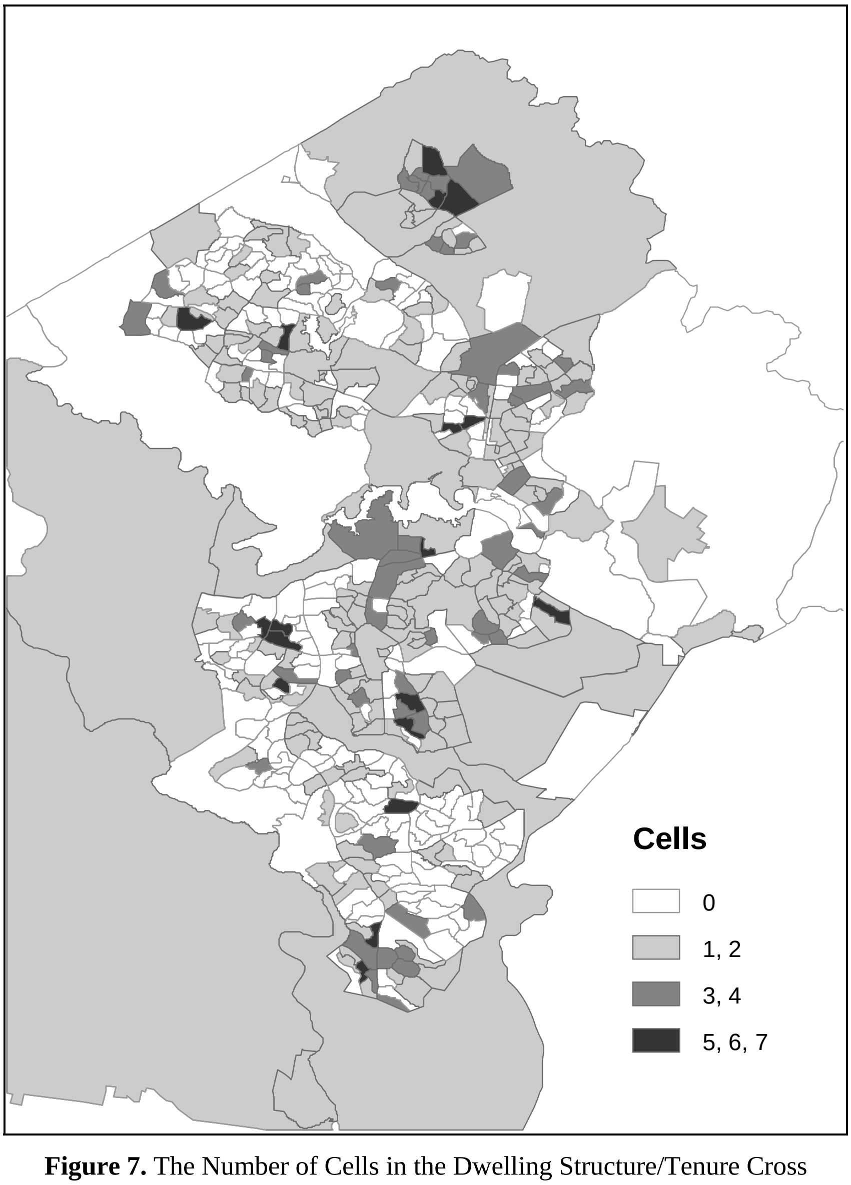


Figure 7. The Number of Cells in the Dwelling Structure/Tenure Cross Tabulation that Differ Significantly from the Census for CCDs in the ACT.

Data Source*:* NATSEM.

determining whether the poor cross-tabulated results for some CCDs is confined to just a few difficult to match variables, such as the tenure of flats, and on methods to improve the optimisation procedure.

Other methods of assessing these cross-tabular results include aggregating CCD level outcomes to a level at which data are available, possibly from other data sources, and purchasing customized Census tables from the ABS against which simulated results can be assessed. These methods will be considered for future validation of SYNAGI results.

Validation of simulated cross-tabulated outcomes is important in that valid correlations between variables and within households is a critical assumption of many potential applications of the SYNAGI method. Relevant customised ABS tables may be available from the ABS to address this issue. This is an issue-specific problem that needs to be addressed in the selection and definition of variables for the particular application being undertaken.

4. OTHER ISSUES

4.1 Socio-demographic Relationships

The underlying assumption in the SYNAGI approach is that there is a strong relationship between socio-demographic profiles of individuals and households in a CCD and the issue or problem of interest. In this way, geographic variations in the socio-demographic characteristics of the population can be used to identify related attributes, preferences, choices and behaviours that vary with these characteristics.

It is likely that this socio-demographic relationship will be stronger for some issues than for others. For example, it is likely that income and family status are highly correlated with expenditure on luxury goods and services, such as overseas holidays and imported motor vehicles. It is less obvious whether these characteristics are related to the length of time taken to travel to work, although there may be a more complex relationship in play for issues such as the choice of work location.

In general terms, this possible socio-demographic relationship can be seen as being either *deterministic*, that is the presence or absence of certain characteristics will determine a particular outcome, or *stochastic,* implying an element of chance or a level of probability that the outcome will occur. Deterministic relationships are conceptually easier to model in that they are rule-based and, although the rules may be complex, the rule can be applied with a level of confidence.

Stochastic relationships on the other hand usually imply preference or choice in behaviour and a probability that a particular household characteristic will affect a certain outcome. Techniques such as Monte-Carlo sampling are then required to produce a geographic distribution of the aggregate outcome based on propensities in the broader population.

4.2 Spatial Effects

Spatial effects refer to the underlying influence that location has on behaviour. Although SYNAGI is spatial, in that it links outcomes to the spatial variation of socio-demographic characteristics in the population, the actual location of the household is not considered in the initial generation of the weights. This can have profound effects on simulated distributions, particularly if factors such as the supply of goods and services, climate and culture are believed to influence the actual outcomes.

These spatial effects can be incorporated into the modelling process by benchmarking the outcomes to known levels. For example, if data are available at state and territory level, expenditures or the incidence of a particular behaviour can be calibrated for differences in a household's state or territory location.

For smaller areas, these spatial effects may be more difficult to adjust for and it is then reliant on the user to interpret the results, incorporating local knowledge or accepting that the results are only indicative.

4.3 Model Development

Although SYNAGI currently uses an optimisation program developed for expenditure applications, the technique can be applied to many other issues. Critical aspects to consider for future applications and development include the selection of integer or fractions of households (that is, whole households or very small weights for many households), sensitivity of the model to the choice and sequence of target variables and the parameters used to control the optimisation process, and the choice of optimisation algorithm. This last point is particularly relevant as a variety of techniques exist that could replace the current algorithm, including genetic algorithms and simulated annealing (Pham and Karaboga, 2000).

It should be borne in mind that synthetic estimation is a developing art that produces data currently unavailable for small geographic areas. For this reason, validation of outcomes is difficult. Many accepted statistical tests and measures, such as standard errors and confidence levels, cannot easily be generated for the SYNAGI model estimates. Indeed, it is questionable whether theses measures are conceptually valid for these new techniques at all, given that the original sample weights are reweighted and factored-down to satisfy a select group of Census targets for each CCD.

As these techniques develop, methods are likely to also develop against which the model outcomes can be assessed. Whether these will be adaptations of existing statistical techniques such as those being developed by Voas and Williamson (2000), only time will tell. For now, validation of SYNAGI relies on internal consistency and comparison with external datasets, and a level of confidence in the conceptual basis of the technique.

5. CONCLUSION

5.1 Assessment of Validation

In conclusion:

- for the majority of CCDs in the ACT, the simulated socio-demographic profiles generated using the weights from the SYNAGI reweighting algorithm match the socio-demographic profiles derived from the Census extremely closely. Overall, the algorithm appears to be achieving the result of reproducing the 64 target variables;
- some targets are less well represented by their simulated values, but only in a

minority of CCDs. This may have implications for some applications of the weights where these targets are particularly relevant to the issue under consideration;

- the results of the cross-tabulation validation exercise, although limited to two variables, suggests that the relationships between simulated variables need to be considered further; and
- although the validation has only looked at the ACT, there is no reason to expect the SYNAGI model to perform differently for other states.

Finally, it should be reiterated that, although this assessment of the convergence process indicates that the reweighting algorithm is achieving a very good result in matching the 64 target variables, by itself this does not mean that these simulated variables can always be recombined to create new multivariate data or that the weights generated can be reliably used with data not included in the optimisation process. These issues will be considered further as the SYNAGI model progresses.

5.2 Future Research and Applications

To date, the SYNAGI methodology has been applied to a variety of issues, including expenditure (Harding et al, 1999), poverty analysis (Lloyd *et al.*, 2001), the delivery of Centrelink services (King *et al.*, 2002) and the 'digital divide' (Hellwig and Lloyd, 2000). Potentially, SYNAGI can be applied to any aggregate survey data that contains sufficient common matching variables with the Census BCPs. The underlying assumption is, as previously discussed, that there is a strong relationship between the socio-demographic profiles of individuals and households within a CCD and the issue or problem being considered. It is likely that the current SYNAGI methodology will work well for some issues and be less useful for others. For this reason, the development of SYNAGI will involve using various methodologies in the selection of appropriate variables, the choice of optimisation algorithm and the application of the weights, to determine the best approach for each application and subject matter.

One of the strengths of using ABS survey data in SYNAGI modelling is that existing microsimulation models that are based on the same ABS surveys can be integrated with SYNAGI. For example, STINMOD - which is NATSEM's tax-transfer microsimulation model and which is widely used by government to assess potential implications of policy change - can be integrated with SYNAGI, so as to provide a picture of the geographic variation in the impact of tax-transfer policy change.

The results to date of the SYNAGI approach are promising, with the optimisation methodology providing an effective method of recreating Census targets for the creation of synthetic microdata for small areas. The challenge now is to assess the robustness of the synthetic microdata, particularly the relationships between the target variables and the possible relationship with unconstrained variables. This level of validation is necessary to consider the potential applications of this method and the tailoring of the SYNAGI model to specific applications.

ACKNOWLEDGMENTS

The authors of this paper would like to acknowledge the earlier work of Dr Otto Hellwig, during the period when he was employed by NATSEM, in the development of the MarketInfo model, which has been the foundation of NATSEM's development of SYNAGI. They would also like to thank Professor Ann Harding, Anthony King and Gillian Beer for their comments on an earlier draft of this paper.

REFERENCES

Australian Bureau of Statistics (1996) *Census Dictionar.y* ABS Cat. No. 2901.0, ABS: Canberra.

Harding, A., Hellwig, O., Bremner, K. and Robinson, M. (1999) *Geodemographics of the Aged: Where They Llive, What They Buy*. Paper presented at the Geodemographics of Ageing in Australia Symposium, Brisbane, 2 December.

Hellwig, O and Lloyd, R. (2000) *Sociodemographic Barriers to Utilisation and Participation in Telecommunications Services and Their Regional Distribution: A Quantitative Analysis.* Report provided to Telstra (available from the NATSEM website at www.natsem.canberra.edu.au).

King, A., McLellan, J. and Lloyd, R. (2002) *Regional Microsimulation for Improved Service Delivery in Australia: Centrelink's CuSP Model.* Paper prepared for the 27th General Conference, International Association for Research in Income and Wealth, Stockholm, Sweden, 18-24 August.

Lloyd, R., Harding, A. and Greenwell, H. (2001) *Worlds Apart: Postcodes with the Highest and Lowest Poverty Rates in Today's Australia*. Paper prepared for the National Social Policy Conference. Sydney, Australia. July.

Pham, D.T. and Karaboga, D. (2000) *Intelligent Optimisation Techniques: Genetic Algorithms, Tabu Search, Simulated Annealing and Neural Networks*. Springer: New York.

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, pp. 349 – 366.

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