



National Centre for Social and Economic Modelling
• University of Canberra •

Comparing Two Methods of Reweighting a Survey File to Small Area Data

- Generalised Regression and Combinatorial Optimisation

Tanton, R , Williamson, P and Harding, A

**Paper Presented to the 1st General Conference of the
International Microsimulation Association
Vienna, 20 - 22 August 2007**

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 with a wide range of Federal and State government agencies.

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.

Director: Ann Harding

© NATSEM, University of Canberra 2007

National Centre for Social and Economic Modelling
University of Canberra ACT 2601 Australia
170 Haydon Drive Bruce ACT 2617

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

Abstract

One method of calculating small area estimates using survey data involves deriving new weights for each respondent in the survey. These new weights are derived so that the survey data sums to some known totals for a small area (from either a Census or administrative data). There are different methods for calculating these weights, and this paper analyses the results from two different methods - a generalised regression method and combinatorial optimisation. The weights derived from each method are compared, and advantages and disadvantages of each method are assessed. Estimates of housing stress at a Statistical Local Area in Australia from each method are then calculated, and these estimates are then validated against a third reliable source, Australian Census data from 2001.

Author note

Robert Tanton is a Principal Research Fellow at the National Centre for Social and Economic Modelling (NATSEM) at the University of Canberra. Dr Paul Williamson is a Senior Lecturer in the Department of Geography at the University of Liverpool. Ann Harding is Professor of Applied Economics and Social Policy at the University of Canberra and Director of NATSEM.

Acknowledgments

The authors would like to gratefully acknowledge the funding support of the Australian Research Council (LP775396) and our partner organisations to the grant - the Queensland Office of Economic and Statistical Research, Queensland Department of Premier and Cabinet, NSW Department of Community Services, ACT Chief Minister's Department and the Australian Bureau of Statistics. The authors would also like to acknowledge earlier preliminary work undertaken on this topic by Shih-Foong Chin.

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.

Contents

Abstract	iii
Author note	iii
Acknowledgments	iii
General caveat	iv
1 Introduction	1
2 Background	2
3 Data and benchmarks	3
4 Differences between the methods	4
4.1 Algorithms	5
4.2 Weights	7
4.3 Efficiency	7
4.4 Summary of differences	8
5 Results from each method	8
6 Further analysis of non-convergent SLAs	13
7 Further analysis of the weights derived by CO and GREGWT	14
8 Conclusions	19
9 Further Work	19
10 References	21

1 Introduction

In 2002, the National Centre for Social and Economic Modelling (NATSEM) started work on a project creating synthetic small area household estimates using Australian Bureau of Statistics (ABS) Survey data and ABS Census data (Chin and Harding, 2006, Chin, *et al.*, 2005). The initial research was supported by Australian Research Council grants (LP0349152 and DP664429), and had Dr Paul Williamson (University of Liverpool) as an international collaborator.

The small area estimation approach adopted by NATSEM involves (re)weighting survey data to agree with a set of known Statistical Local Area totals drawn from ABS Census data. This is directly analogous to the more conventional situation in which survey weights are inflated to fit a known set of State or national totals - the main difference being that, to fit small area totals, survey weights typically have to be deflated rather than inflated.

For the purposes of small area estimation two methods of reweighting survey data have been explored. The first uses an iterative Generalised Regression algorithm which attempts to minimise a truncated exponential distance function. This algorithm is implemented using a program developed by the Australian Bureau of Statistics called GREGWT (Bell, 2000). This program is used by the ABS to benchmark their survey data to known State totals.

The second method considered uses an iterative 'combinatorial optimisation' algorithm, in which an initial set (combination) of households are drawn from a survey at random (with replacement), following which a succession of random changes in the households selected are made, with a view to optimising the fit of the household combination to the specified small area benchmarks. This algorithm is implemented using the program CO, developed by Dr Paul Williamson at the University of Liverpool (Williamson, 2007, Williamson, *et al.*, 1998).

Unless earlier exit criteria are met (convergence measures for GREGWT; estimate fit thresholds for CO), both algorithms continue until a maximum number of user-specified iterations has been exceeded.

This paper compares the two algorithms in terms of their advantages and disadvantages, including the number of SLAs which fail to satisfy minimum fit criteria, and the resulting weights from each method.

Section two of this paper outlines the data and benchmarks used for both methods. The data and benchmarks used for each method are exactly the same. Section three summarises the differences between the two methods, in terms of the methods and assumptions. Section four compares the results from each method, looking at the

total difference and differences for each benchmark. Section five provides some analysis of the non-converging SLAs. Section 6 provides further analysis of the weights - including looking at the distribution of the weights from each procedure - while Section 7 provides conclusions and directions for further work.

2 Background

There has been considerable work in Australia and Britain on generating small area estimates using survey data. The attraction of small area models is that they allow a survey designed for generating reliable estimates for a large area to be used to derive reliable estimates for a small area, without increasing the sample size, which is an expensive process.

The Australian Bureau of Statistics recently produced a small area estimation practice manual (ABS, 2006), which outlines some of the techniques, and includes a section on diagnostics and quality measurement. The manual covers simple small area methods, like broad Area Ratio Estimator and Calibration estimators; and then covers regression methods, including random effects regression models. While this manual is theoretical, the ABS has produced a number of small area estimates using a variety of techniques.

In 2005, the ABS produced small area estimates of disability (ABS, 2005), which looked at three different methods of estimating disability for small areas; a Poisson regression model; a Bernoulli model; and ratio estimation. The report found that the Bernoulli model and the ratio estimator gave the best results, with the Poisson model performing poorly, possibly due to overdispersion.

The ABS also used a number of methods to generate small area estimates of crime. While some of this was unpublished, a method using a regression estimator was published (Tanton, *et al.*, 2001). The results from a number of methods, including a ratio method based on an article by Purcell and Lincare (Purcell and Linacre, 1976) and a Structure Preserving Regression Estimator (SPREE) used for estimating labour force by the ABS, were unpublished, due to the fact that the results were difficult to validate. This has been the biggest difficulty with the small area estimators derived by the ABS - there is no estimate of the reliability of the results, for example, standard errors or confidence intervals.

Outside of the ABS, the Commonwealth Department of Employment and Workplace Relations uses a SPREE approach to estimate small area labour force statistics (Commonwealth Department of Employment and Workplace Relations, 2007). The

ABS is now re-examining the estimation of the labour force using small area estimation techniques (ABS, 2007).

The National Centre for Social and Economic Modelling has also produced small area estimates of poverty and housing stress, using the methods outlined in this paper. These estimates were published in 2006 (Chin, *et al.*, 2006, Harding, *et al.*, 2006, Phillips, *et al.*, 2006).

In the UK the Office for National Statistics has conducted a review of alternative methods for updating small area population estimates between censuses, in lieu of a population register, concluding by favouring a ratio change approach (Bates, 2006). An allied initiative has seen the development of small area estimates using a multilevel regression-based synthetic estimator fitted using area-level covariates. The result has been the release of a series of 'experimental' small-area statistics covering topics such as income, household overcrowding and social capital (Heady, *et al.*, 2003). Both articles can be downloaded from the ONS website at www.ons.gov.uk

3 Data and benchmarks

Both the methods of regional microsimulation being compared here require two sets of data. One set is the survey which is being reweighted; and the second is the set of benchmarks that the survey is being reweighted to. The benchmarks must be reliable for the small area being estimated.

In this case, we have used data from the 1998-99 Household Expenditure Survey from the ABS. This is a confidentialised unit record file, which we then manipulate by adding a record for each child (the CURF only has the total number of children in the household) and for each person living in a non-private dwelling (which are not on the CURF but are on the data being benchmarked to).

The benchmarks for the reweighting process come from the 2001 Census. The Census provides reliable data for SLAs, and the Expanded Community Profile tables provide the cross-tabulations that we require. Unfortunately the raw Census data include 'Not Stated' counts. These are counts of people or households that did not respond to certain questions on the Census. This can be partial non-response (eg, they said they were employed, but did not say whether this was full time or part time); or full non-response (eg, they didn't answer the employment question). The 'Not Stated' values are distributed by us across all valid responses, using an integer pro-rata method in which any unit remainder is allocated to the category with the highest value.

In previous work, we have also 'balanced' the data. This involves adjusting the values of cells until the cell totals and sub-totals in each table match. The reason they may not match is due to the ABS randomising small cell counts. Experimenting with balanced and unbalanced data has shown that balancing has little effect on results, so the original unbalanced census counts were used for this project.

Further information on how the survey and Census data are adjusted can be found in one of NATSEM's technical papers, available from the NATSEM website (Chin, *et al.*, 2006).

From the survey and Census data, we have chosen a set of benchmark variables that are available on both sets of data, aggregating variable categories in one or other set of data until the categories are exactly the same. The final set of benchmarks used for this project are shown in Table 1.

Table 1 **Benchmarks used for creating small-area weights**

Census Table	Type	Dimension ¹	Fully specified ²	Benchmarks (no.)
Age by sex by labour force status	Person	Multi	Yes	32
Residents in different types of non-private dwelling	Person	Single	Yes	8
Household Type	Household	Single	No	1
Household size - number usual residents	Household	Single	Yes	7
Dwelling tenure by weekly household rent	Household	Multi	No	7
Dwelling tenure by household type	Household	Multi	Yes	15
Dwelling tenure by weekly household income	Household	Multi	No	16
Monthly household mortgage by wkly household income	Household	Multi	Yes	12
Weekly household rental by weekly household income	Household	Multi	Yes	20
Dwelling structure by household family composition	Household	Multi	No	12
Total number of benchmark tabulations				10
Total number of benchmarks				130

1. Multi-dimensional means cross-tabulations of variables.

2. Not fully specified means that one or more of the cells in a benchmark tabulation were not used for weight production. For example, for the benchmark table of 'Household Type', the count of 'Private households' was extracted for use as a benchmark, whilst the count of 'Non-private dwellings' was excluded from the reweighting process

4 Differences between the methods

There are a number of theoretical differences between the two methods that need to be outlined.

4.1 Algorithms

The algorithms used for each method are quite different. The GREGWT algorithm is essentially a constrained distance minimisation function. The method uses regression to get an initial weight; and then iterates the regression until convergence is reached (so the difference between the estimated benchmark and the actual benchmark for the area from the Census data is within a set limit), or a set number of iterations is made, at which time the iteration stops. The process needs a start weight, and this is set to the original ABS survey weight for the survey record divided by the population of the SLA. In many cases, there is no iteration as the initial regression estimate provides weights that are within the tolerance set.

Full information on the GREGWT macro can be found in the user manual for GREGWT (Bell, 2000). As implemented for this paper, the option to minimise change in individual household weights through using a truncated exponential distance function was turned on.

At the end of the reweighting process, every household in the survey dataset will have a weight for each Census SLA for which benchmark counts were supplied. In a small number of cases these weights will have been generated even though estimate convergence was not achieved, due to the algorithm halting after exceeding a user-specified number of iterations (30 for the work reported in this paper). In these non-convergent cases the weights from the terminal iteration typically, but not always, include a small number of exceptionally high household weights, leading to very poor fits to one or more benchmark counts. We have chosen not to discard all non-convergent GREGWT outputs, as a few technically non-convergent estimates actually give rise to sets of weights that fit all benchmarks reasonably well. Instead, for the purposes of this paper, we identify as ‘non-convergent’ any SLA for which the sum of the absolute value of all errors across all benchmarks is greater than the number of households in the SLA; so where:

$$\frac{\sum_{\text{Benchmark}} \text{ABS}(\text{Actual} - \text{Estimated})}{N_{hh}} > 1$$

The Combinatorial Optimisation algorithm, as currently implemented, may be viewed as an integer reweighting algorithm. For each household in an SLA (as recorded in the Census benchmark counts), CO randomly selects one household (with replacement) from the survey dataset. This is equivalent to setting all survey household weights to 0, then incrementing household weights at random by a count of 1 until the sum of weighted households matches the equivalent benchmark count. In each subsequent iteration the weight of one household is randomly increased by one, whilst the weight of another (non-zero weighted) household is randomly

decreased by one. This is equivalent to randomly swapping households in and out of the set (combination) of households currently selected to represent the SLA. If the change in weights leads to improved fit, the change is retained; a few adverse changes in weights are also accepted, with a probability that diminishes in proportion to (i) size of the adverse impact and (ii) number of iterations, in order to avoid getting stuck in a local sub-optima; otherwise the change is rejected and the weights are reverted to their previous values. CO will continue to iterate until either a minimum fit threshold is achieved, or until a maximum number of user-specified iterations has been exceeded (5 million for the results reported in this paper). Full details of the algorithm and its links to simulated annealing are published in an article by Williamson *et al* (Williamson, *et al.*, 1998). For details of the latest publicly available version of the CO program, the user manual has recently been published by Paul Williamson (Williamson, 2007).

When deciding whether or not to accept a change in household weights, CO can evaluate one of two measures of fit. The first is the Overall Total Absolute Error. This is a conventional measure of fit that seeks to minimise the absolute difference between benchmark counts and their weighted survey equivalents, and is given by

$$OTAE = \sum_{ij} o_{ij} - e_{ij}$$

where e_{ij} is the expected (census) count for cell j in benchmark table i and o_{ij} is its estimated (weighted survey) equivalent.

The second measure is the Overall Relative Sum of Squared modified Z-Scores (ORSSZm²), defined as

$$ORSSZm^2 = \sum_i \frac{1}{c_i} \left[\sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij} \left(1 - \frac{e_{ij}}{\sum_j e_{ij}} \right)} \right]$$

where c_i = the χ^2 critical value for benchmark table i , with $p=0.05$ and $d.f.$ = number of cells in table. The derivation of this second measure is explained in full in an article by Voas and Williamson (Voas and Williamson, 2001). The underlying principle behind this second measure is the use of a modified Z-score for each benchmark count that takes into account not only the proportional difference between observed and expected counts (as with a normal Z-score), but also the absolute difference between estimated and observed benchmark table totals. The resulting sum of squared modified Z-scores for each benchmark table is divided by the relevant table-specific χ^2 critical value to standardise for the number of benchmark counts in each benchmark table. These table-specific relative scores are

then summed to yield the overall measure, the main focus of which is upon proportional rather than absolute fit.

Whereas the results presented in this paper mainly exclude 'non-convergent' (i.e. very poorly fitting) GREGWT estimates, they include all CO estimates, whether or not these estimates satisfied the minimum fit thresholds specified for triggering early termination of a CO run on the grounds that non-convergence, *per se*, is not an issue for CO.

4.2 Weights

Because of the methods used, GREGWT produces floating point weights - whereas CO produces integer weights. There is no real advantage to either type of weight. In fact, the CO routine could implement floating point weights by adding partial 'units' of individuals, rather than whole records.

4.3 Efficiency

Both the CO and GREGWT routine are computationally intensive. In testing the different algorithms on exactly the same computer (dual processor dual core 2 Ghz processor, 2 GB Memory), the CO routine calculated weights for the 107 SLAs in the Australian Capital Territory in about ½ hour; where the GREGWT routine took 2 ½ hours. This was possibly due to the way the algorithms are coded (GREGWT is in a SAS macro; whereas CO uses compiled FORTRAN code), but may also reflect algorithmic efficiency.

4.4 Summary of differences

A summary of the differences in each method is shown in Table 2.

Table 2 **Comparison of methods in summary**

	GWT	CO
Approach	National household weights from a national survey dataset are reweighted to household weights for SLAs by constraining to small-area census counts	Selection of a combination of households from a national survey microdata set that best fit small-area census counts
Weights	In fractional numbers	In integer numbers
Preparation of census data	Needs to address re-allocation of 'not-stated' and 'not applicable' counts	Needs to address re-allocation of 'not-stated' and 'not applicable' counts
Optimisation strategy	Algorithm reaches an optimised solution when residual (i.e. difference between an synthetic estimate and the benchmark count) approaches zero	Minimise absolute or proportional error
'Convergent' & 'non-convergent' SLAs	In some cases no convergent solution may be found; Average Household Absolute Sum of Residuals is ≤ 1 provides a proxy indicator for this non-convergence.	No convergence issues, although final 'optimal' estimates may still fail to fit all user-supplied benchmarks.

Source: NATSEM (GWT) and Williamson (CO)

5 Results from each method

This section shows summary results for each method. The first set of results shows how well each method has hit the benchmarks specified. The second set of results, more interestingly, show the usefulness of each method for predicting values not present in the benchmarks.

Predicting the benchmarks (constrained variables) is of limited interest, as by definition we already know their values. On the other hand, the difference between the estimated (weighted survey) and observed (census) values does at least provide some initial indicator of estimate quality. More usefully, the weighted data also yield estimates for unbenchmarked values. Two kinds of value-added estimates may be identified. The first involves the unbenchmarked interactions between benchmarked variables (margin constrained estimates). The second involves the unbenchmarked interactions between unbenchmarked variables (unconstrained estimates). The greater the degree of correlation between the benchmark constraints and the unbenchmarked estimates, the greater the quality of the estimate is likely to be. In contrast, values estimated using unbenchmarked variables that have no correlation with the benchmarked variables will necessarily be highly unreliable.

In this paper we evaluate the efficacy of our small area estimates in predicting housing stress. Housing stress is directly correlated with three of our constrained variables: income, rent paid and mortgage paid. A household is defined here as being in housing stress when they spend more than 30 per cent of their gross income on rent or a housing loan (a definition that can be matched by the ABS from the Census data).

The average SLA-level fit to the benchmarks listed in Table 1 is summarised in Tables 4 and 5 for two States in Australia, New South Wales and the Australian Capital Territory, using a range of summary statistical measures described in Table 3. There are 107 areas estimated in the Australian Capital Territory and 199 in New South Wales. The SLAs in ACT are atypical for Australia, both in terms of socio-demographic composition and in terms of size (ACT SLAs contain considerably smaller populations than average). The SLAs in NSW may be regarded as more 'typical' in both regards.

Table 3 Summary measures of goodness of fit

Measure	Description
Overall Total Absolute Error (OTAE)	Absolute Sum of Residuals summed across all benchmark counts
Overall Total Absolute Error per household (OTAE/HH)	Absolute sum of residuals per household across all benchmark counts
Overall Total Absolute Proportional Error (OTAPE)	Absolute difference between benchmark counts when expressed as fraction of the table total
Overall relative sum of Z-square scores (ORSumZ2)	For each benchmark table, the Z-score of each benchmark count squared, and summed for the table; then divide by chi-square critical value for table (--> RSumZ2), then sum across all tables (-->ORSumZ2). For a given table, RSumZ2>1 shows it is not fitting.

The results are shown in Table 4 and Table 5, averaged across the SLAs for which GREGWT produced 'convergent' estimates.

Both variants of CO produced a better proportional fit to the estimation benchmarks than GREGWT (lower ORSumZ2 and OTAPE) but performed more variably when it came to matching GREGWT's absolute fit to the estimation benchmarks. Of the two CO variants, CO (Min Proportion) unsurprisingly produced by far the lowest proportional error for both States; more surprisingly it also produced the lowest absolute error in NSW. This may reflect the fact that the SLAs in NSW are more populous, making the link between absolute and proportional values more tenuous. Overall the evidence presented in Tables 4 and 5 suggests a performance advantage favouring selection of CO (Min Proportion) over CO (Min Absolute), if such a choice has to be made. For non-covergent SLAs (results not presented here), both variants of

CO significantly out-performed GREGWT on all measures (on account of the extremely high weightings given by GREGWT to a few households).

Table 4 Results for constrained variables, Australian Capital Territory (GREGWT 'convergent' SLAs only)

Measure	GREGWT	CO (Min Proportion)	CO (Min Absolute)
OTAE	139.6	133.4	92.2
OTAE/HH	0.1	0.1	0.1
OTAPE	0.4	0.2	0.2
ORSumZ2	48.4	0.5	27.8

Note: Lower numbers signify greater accuracy

Table 5 Results for constrained variables, New South Wales (GREGWT 'convergent' SLAs only)

Measure	GREGWT	CO (Min Proportion)	CO (Min Absolute)
OTAE	602.9	483.1	979.3
OTAE/HH	0.1	0.1	0.1
OTAPE	0.2	0.1	0.2
ORSumZ2	60.5	1.9	29.2

Note: Lower numbers signify greater accuracy

As mentioned above, the main use for these reweighting techniques is to get estimates for variables that were not on the Census. We should be able to get reasonable estimates for variables that are correlated with the benchmarked variables (margin-constrained variables). For this paper, we calculated estimates of housing stress, which are correlated with some of the benchmarked variables (Income and Housing costs). Estimates of housing stress supplied by the Australian Bureau of Statistics from Census data provide an independent source against which to compare our own estimates.

A comparison of our various modelled estimates with those supplied by the ABS is presented in Table 6 for the Australian Capital Territory and New South Wales. It can be seen that all the models produced reasonable estimates of proportions, with the GREGWT estimates coming closest to the actual ABS per cent figures. However, this is after removing all the non-convergent SLAs.

There were 24 non converging areas in the Australian Capital Territory, and 13 in New South Wales. Further analysis, and a list of non converging areas, is presented in Section 6 of this paper.

Table 6 Results for Housing Stress, Australian Capital Territory and New South Wales (GREGWT 'convergent' SLAs only)

State	Number Unaffordable			
	ABS	GREGWT	CO (Min Proportion)	CO (Min Absolute)
Australian Capital Territory	5,526	6,147	5,924	5,821
New South Wales	169,823	194,394	191,720	189,269
Total	175,349	200,541	197,644	195,090
	% Unaffordable			
Australian Capital Territory	5.9	5.9	5.7	5.6
New South Wales	9.1	9.2	9.1	8.9
Combined	9.0	9.0	8.9	8.8

In terms of the number of people in housing stress, Table 6 appears to suggest that neither GREGWT nor CO estimate absolute numbers all that well. However, in considering Table 6 it should be borne in mind that the population base for the independent ABS housing stress estimates includes only households providing full returns on income and housing costs via their Census form. In contrast, the GREGWT and CO estimates have been weighted to fit benchmarks in which non-response households have been included via pro rating (see section 3). Given that pro rating should more or less preserve the proportional distribution of households by income and housing cost, and given that the CO and GREGWT estimates closely replicate ABS estimates of the proportion of households in housing stress, it could in fact be argued that it is actually the ABS absolute estimates that are out of line. In any case, it is certainly true that GREGWT and CO outputs do need to be recognised as modelled estimates. As such, estimated proportions are more likely to be accurate than estimated levels - and so results are best presented as proportions or grouped into quantiles.

Another way of looking at these results is to look at the correlation between the different methods for each SLA in the area. The graphs below show the correlations between the ABS estimate and the different estimation methods we have used for all convergent SLAs in the Australian Capital Territory and New South Wales. It can be seen that the correlations are all very high (0.86 - 0.89). The highest correlation (and therefore lowest error) is using the CO-Min Proportion model. Clearly the ranking of the three model estimates depends upon the precise measure used. But overall all three approaches appear to have done a good job of estimating the unknown three-way interaction between income and tenure-specific housing cost, from which the final estimate of housing stress is derived.

Figure 1 **SLA level Housing stress estimates (%): GREGWT (GREGWT convergent SLAs)**

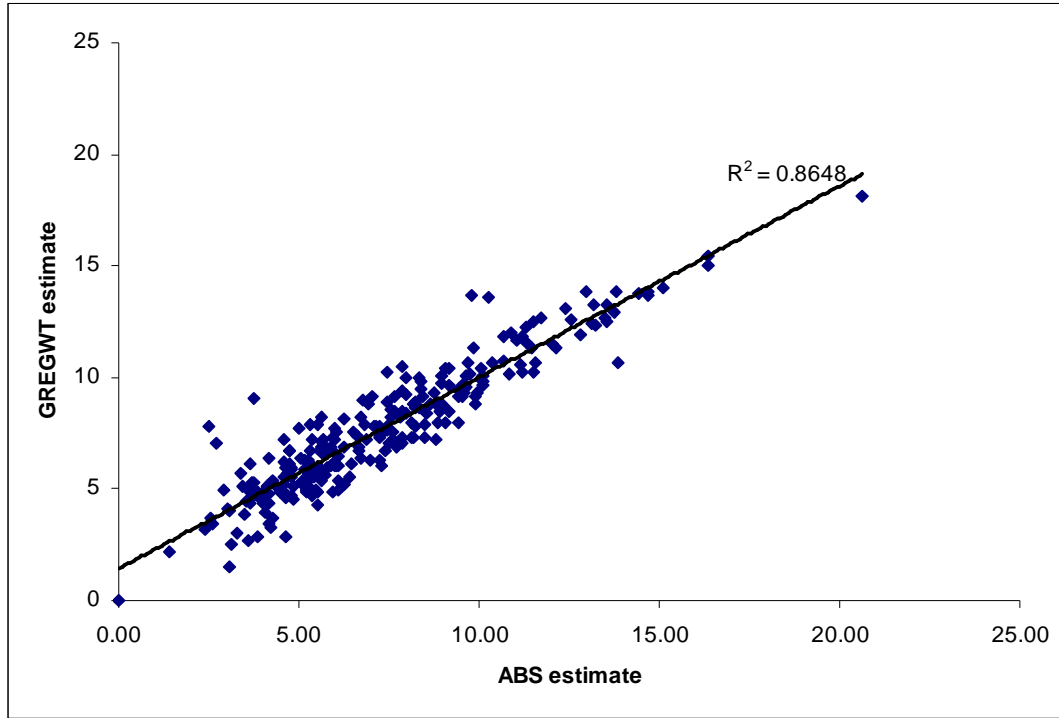


Figure 2 **SLA level Housing stress estimates (%): CO-Min Proportion (GREGWT convergent SLAs)**

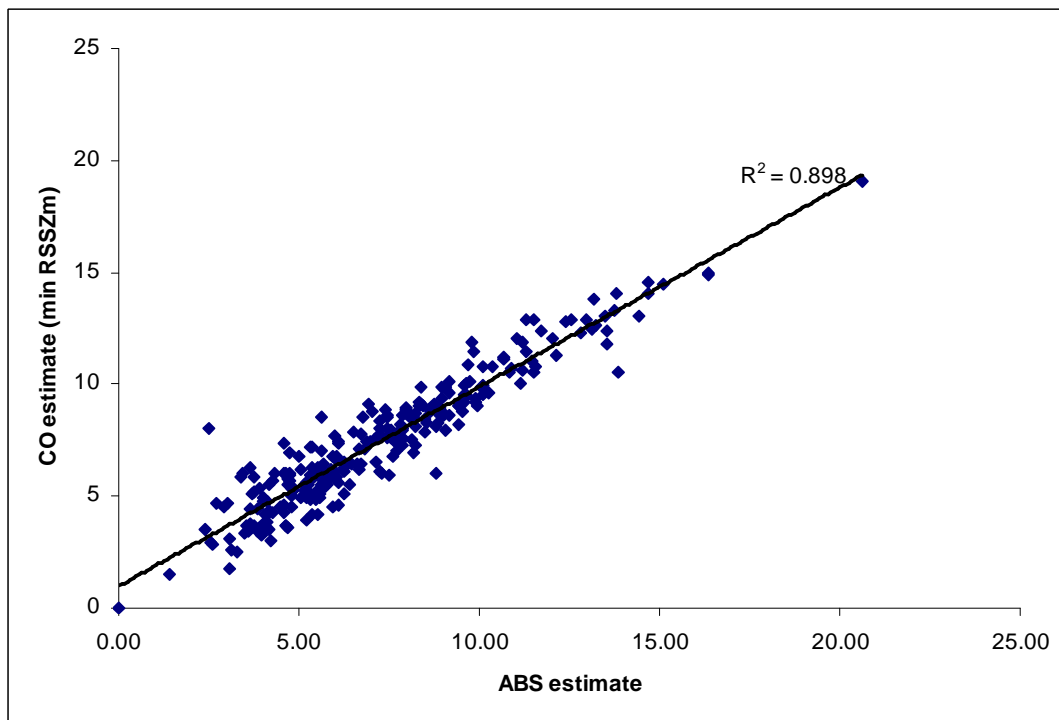
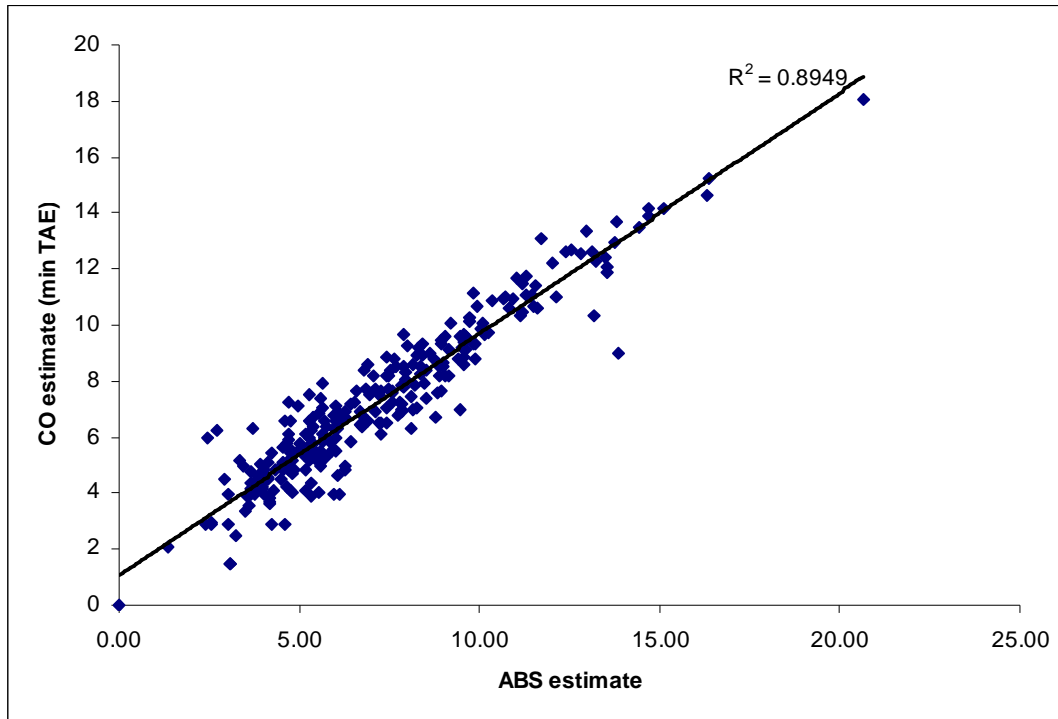


Figure 3 **SLA level Housing stress estimates (%): CO-Min Absolute (GREGWT convergent SLAs)**



6 Further analysis of non-convergent SLAs

As already noted, GREGWT is an iterative algorithm. The procedure will iterate until either a set of weights are found that satisfy all of the user-supplied benchmarks, or a set limit has been reached, at which point the area is deemed to be non-convergent. The maximum number of iterations is set in the GREGWT code, and for these simulations was set to 30. Experimentation has shown that additional iterations lead to only very marginal improvements in estimates at the cost of considerably increased compute times.

Most, but not all, cases of non-convergence are associated with poor levels of fit between weights and benchmarks. In this section we turn our attention to a brief analysis of those 'non-convergent' SLAs for which the absolute value of the sum of difference between each benchmark and its weighted estimate, divided by the number of households in the area, was greater than one. Previous experience has

identified this as an upper threshold beyond which the output weights produce estimates of no practical value.

The number of non-converging areas for this test was 24 in the Australian Capital Territory and 13 in New South Wales, giving a total of 37 areas out of 306 - so about 12 per cent of areas were lost due to non convergence. Non-convergence affected the Australian Capital Territory the worst, with 22 per cent of areas non-convergent.

Many of these non-converging SLAs have low population, so a more interesting statistic is the number of households in non-converging SLAs. In the New South Wales, there were 131,155 households in non-converging SLAs, representing about 5% of the total number of households. In the Australian Capital Territory, there were 8,836 households in non converging SLAs out of a total of 121,265 households, so about 7% of households were lost due to non-convergence.

The GREGWT algorithm could not produce numbers for these SLAs, whereas the CO algorithm was able to produce numbers which relatively closely matched both the initial weighting benchmarks and the ABS housing stress estimates. Many of the non-convergent SLAs are areas with very low populations; or areas that are different to other SLAs in the state. For instance, they may have a high proportion of retail space rather than residential, and the residential space may be a mix of public housing and expensive units (for instance, inner cities); or they may have a high transient population (for instance, industrial areas). Some of these areas may include remote areas, off-shore islands, industrial estates and inner city areas experiencing urban renewal. For many users, these atypical areas are not the interesting areas, so not having estimates is not seen as a major problem.

7 Further analysis of the weights derived by CO and GREGWT

While Section 5 has shown that the results from estimating different variables using these two methods are similar, the weights derived by each method are quite different. The CO routine derives integer weights - whereas the GREGWT routine doesn't. We also expect more zero weights from the CO routine than we get using GREGWT. It would be interesting to look at the distribution of these weights.

The number of weights derived by each routine is massive. There are 6,892 households on our survey file that we derive weights for. In the Australian Capital Territory, there are 107 SLAs. So in the dataset of weights for the Australian Capital Territory, there are a total of about 740,000 weights calculated. For New South Wales, with 199 areas, there are about 1.4 million weights calculated.

Table 7 shows the size distribution of the weights from CO and GREGWT. The CO routine only uses integer weights, so if there are fewer households in an SLA than in a survey, there will inevitably be survey households with a weight of 0. In contrast, GREGWT shares out the weights in small fractions across a large number of households. It can be seen, therefore, that the CO routine produces many more 0 weights than GREGWT.

Table 7 **Size of weights from CO and GREGWT**

Method	Australian Capital Territory			New South Wales		
	0	>0 – 1 ^a	>1	0	>0 – 1 ^a	>1
	%	%	%	%	%	%
CO	95	3	1	79	8	13
GREGWT	53	47	1	36	48	16

^a For >0 – 1, the CO value is 1.

Source: NATSEM modelled data

This also means that the CO routine is relying on fewer households to calculate the values of housing stress in the previous example. The GREGWT routine will use more households, with lower weights. The distribution of the weights from GREGWT for the Australian Capital Territory and New South Wales is shown in Figure 4 and Figure 5. Both these frequency distributions have class boundaries increasing by 0.01 until the value of one is reached, and the final category is weights greater than one. It can be seen that most of the weights in the Australian Capital Territory are below one (99 per cent of them, according to Table 7); in New South Wales, 15 per cent are above one, but of these, half of them are under 2.

Frequency distributions have not been shown for the CO routine, as 95 per cent of the weights are 0 for each State, so the frequency distributions are dominated by this.

The other interesting statistic to look at with the weights is the maximum and average weights. For the GREGWT routine, non-convergence can lead to weights that are ridiculously large (in the order of 10^4 or more). Excluding SLAs for which GREGWT produced non-convergent estimates, the maximum and average weights produced by GREGWT and CO are shown in Table 8. It can be seen that the CO routine produces a higher maximum for the Australian Capital Territory - but GREGWT produces a higher maximum for New South Wales.

The average values shown in Table 8 are calculated without zeroes, as there are so many in each procedure that they dominate an average. The values in Table 8 show that the CO routine calculates higher weights on average than the GREGWT routine for both New South Wales and the Australian Capital Territory. This is to be expected, given the number of zero weights calculated by the CO routine. To get to the same population in an area, the positive weights must be higher in the CO routine than the GREGWT routine, which has fewer zero weights.

Figure 4 **Distribution of GREGWT weights for NSW**

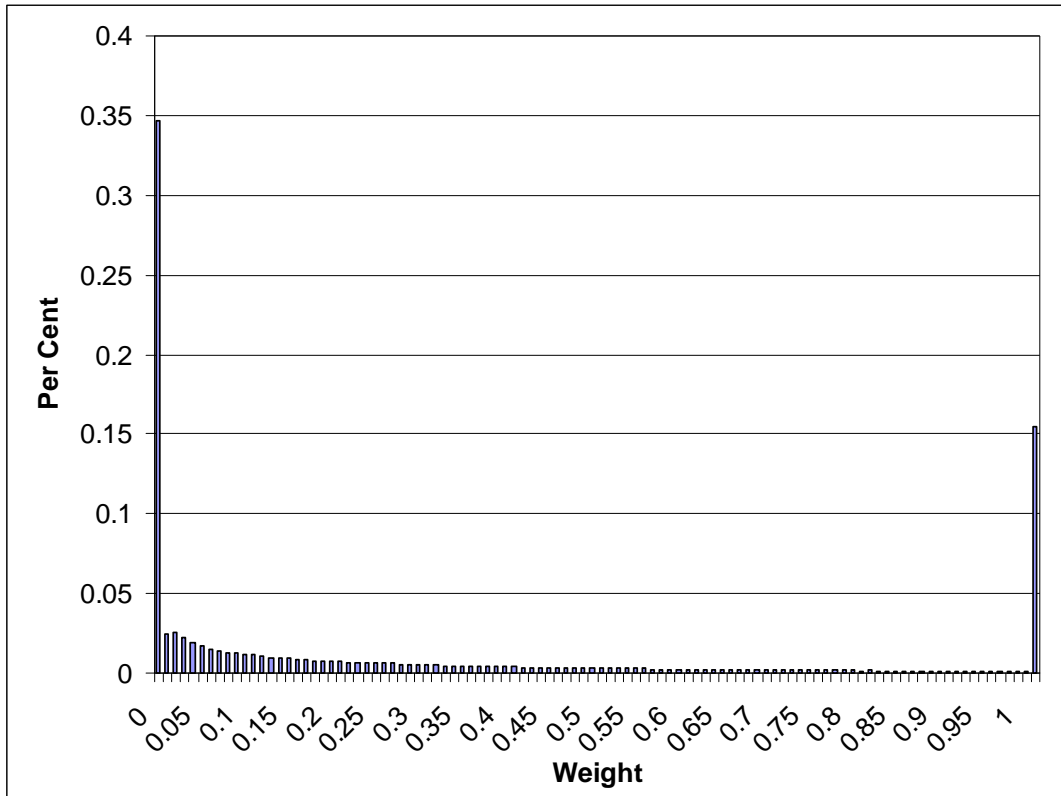


Figure 5 Distribution of GREGWT weights for the Australian Capital Territory

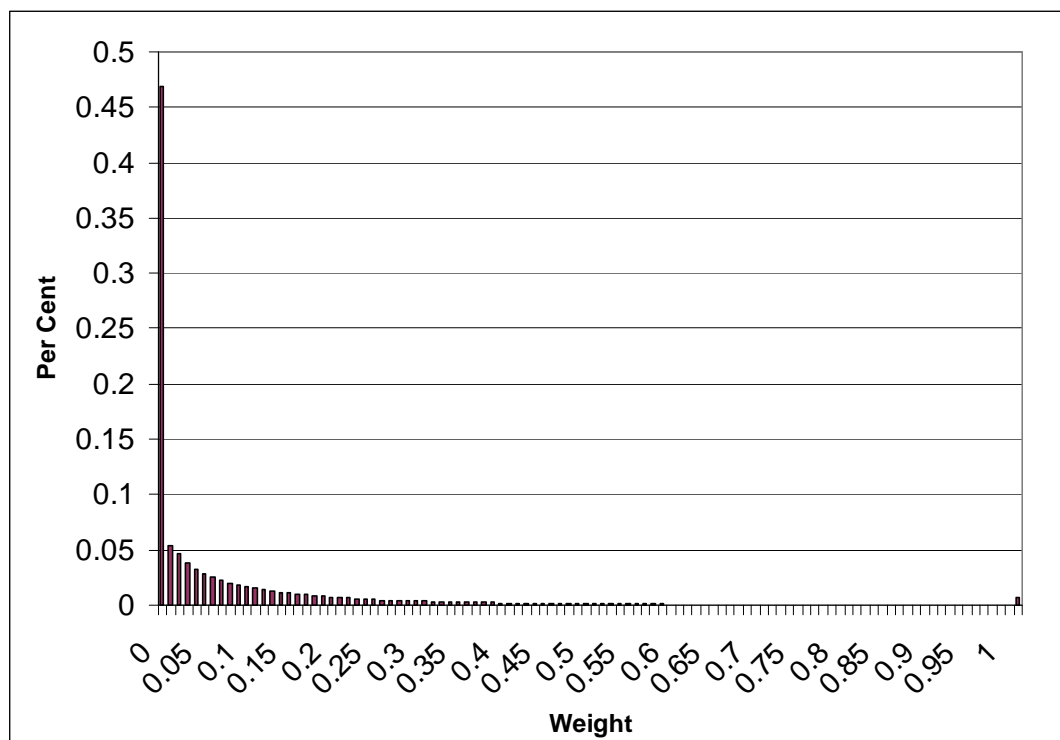


Table 8 Weights for GREGWT and CO (GREGWT convergent SLAs only)

Method	Maximum		Average non-zero value	
	New South Wales	Australian Capital Territory	New South Wales Average	Australian Capital Territory Average
CO (Min Proportion)	443	24	3.49	1.45
GREGWT	647	18	1.11	0.15

This section has shown that even though the weights from the CO and GREGWT algorithm give similar results when calculating variables like housing stress, they are actually very different. The CO routine tends to include fewer households, but give them higher weights – while the GREGWT routine will select more households to represent an SLA, but will give them smaller weights.

8 Conclusions

The GREGWT algorithm has been shown to be capable of producing good results, but for 14% of the total number of SLAs did not converge. This means that a significant minority of areas have no usable estimate, a significant limitation of the GREGWT algorithm. The GREGWT algorithm also takes a long time to run compared to CO, particularly when estimating a large number of areas. It is unknown whether this is due to the programming language being used for each algorithm (CO uses a compiled FORTRAN code, whereas GREGWT is a macro running in SAS) or whether this reflects the relative efficiencies of the underlying algorithms.

Head-to-head, when results are compared for those SLAs for which GREGWT converged, the fit to benchmarks and estimates of housing stress produced by GREGWT and CO (Min Proportion) are broadly comparable. On balance, however, the CO Min Proportion model is perhaps slightly to be favoured. It has a better proportional fit to benchmarks (Table 4), the lowest error when estimating each SLA's value (Figure 2), and produces reasonable estimates when the SLA values were aggregated to State (see Table 6). In addition, for those SLAs for which GREGWT produces no usable estimate, CO (Min Proportion) appears to continue to produce estimates of reasonable quality. CO (Min Absolute) competes slightly less well head to head with GREGWT, and in almost all circumstances produces estimates of at least marginally lesser quality than those offered by CO (Min Proportion).

9 Further Work

One of the questions we have not looked at is how complex the benchmarks can be for each method. One of the limitations of the reweighting process is that as the number of benchmarks increases, the difficulty of finding a set of weights that simultaneously satisfies all of benchmarks also increases, leading at minimum to increased model run times. Experience to date suggests that increasing numbers of benchmarks impacts more adversely on GREGWT than on CO, partly in terms of run time, but particularly in terms of significantly increasing levels of non-convergence. One of the things we are planning to do next is to run both models using more benchmarks, drawn in particular from more complex multivariate distributions.

There is also further work being carried out bringing new survey data in, more than doubling the survey households available for weighting by switching to a new survey and pooling observations across multiple adjacent years. This work will have

the added benefit of enabling NATSEM to link the weights from the regional model directly to data from our STINMOD microsimulation model, as the survey data will be exactly the same as the survey data used for STINMOD. This will also mean any policy simulations done using STINMOD can be 'regionalised', assuming the benchmarks are appropriate.

A final area to consider is the possibility of converting CO from an integer to a fractional weighting algorithm, in which each iteration involves considering the adjustment of two household weights by complementary fractions of, say, +0.1 and -0.1. This is likely to lead to longer run times, but may help to improve estimates by avoiding high weights for an overly narrow selection of survey households.

10 References

- ABS (2005) *Small area estimates of Disability in Australia*, Canberra: ABS, Pub. # 1351.0.55.006
- ABS (2006) *A guide to small area estimation - Version 1.0*, [http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/3a60738d0abdf98cca2571ab00242664/\\$FILE/May%2006.pdf](http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/3a60738d0abdf98cca2571ab00242664/$FILE/May%2006.pdf), Last accessed 30 July 2007
- ABS (2007) *Small area estimation of LFS*, <http://www.abs.gov.au/AUSSTATS/abs@.nsf/7d12b0f6763c78caca257061001cc588/a33405ed6992b6fdca25730f0018fa33!OpenDocument>, Last accessed 30 July 2007
- Bates, A (2006) 'Methodology used for producing ONS's small area population estimates', *Population Trends*, 125(30 - 36)
- Bell, P (2000) *GREGWT and TABLE macros - Users guide*, Canberra: ABS, Pub. # Unpublished
- Chin, S-F and Harding, A (2006) *Regional Dimensions: Creating synthetic small-area microdata and spatial microsimulation models*, Technical Paper 33, NATSEM
- Chin, S-F, Harding, A and Bill, A (2006) *Regional dimensions: Preparation of 1998-99 HES for reweighting to small area benchmarks*, Technical Paper 34, NATSEM
- Chin, S-F, Harding, A, Lloyd, R, McNamara, J, Phillips, B and Vu, Q (2005) 'Spatial microsimulation using synthetic small-area estimates of income, tax and social security benefits', *Australasian Journal of Regional Studies*, 11(3), 303 - 336
- Chin, S-F, Harding, A and Tanton, R (2006) 'A spatial portrait of disadvantage: Income poverty by Statistical Local Area in 2001', Paper given at 2006 ANZRSI Conference Heritage and regional development, Beechworth, Victoria, 26 - 29 September 2006
- Commonwealth Department of Employment and Workplace Relations (2007) *Small Area Labour Markets*, Commonwealth Government
- Harding, A, Lloyd, R, Bill, A and King, A (2006) *Assessing poverty and inequality at a detailed regional level: New advances in spatial microsimulation*, Helsinki: United Nations University Press
- Heady, P, PClarke, P, Brown, G, Ellis, K, Heasman, D, Hennell, S, Longhurst, J and Mitchell, B (2003) *Model-Based Small Area Estimation Series No. 2 - Small Area Estimation Project Report*, Office of National Statistics
- Phillips, B, Chin, SF and Harding, A (2006) 'Housing Stress Today: Estimates for Statistical Local Areas in 2005', Paper given at Australian Consortium for Social and Political Research Incorporated Conference, Sydney, 10-13 December 2006
- Purcell, N and Linacre, S (1976) 'Techniques for the Estimation of Small Area Characteristics', Paper given at Third Australian Statistical Conference, 18 - 20 August 1976
- Tanton, R, Jones, R and Lubulwa, G (2001) 'Analyses of the 1998 Australian National Crime and Safety Survey', Paper given at The Character, Impact and Prevention of Crime in Regional Australia, Townsville, 2 - 3 August 2001
- Voas, D and Williamson, P (2001) 'Evaluating goodness-of-fit measures for synthetic microdata', *Geographical and Environmental Modelling*, 5(2), 177 - 200

Williamson, P (2007) *CO Instruction Manual*, Working Paper 2007/1, Population Microdata Unit, Dept. of Geography, University of Liverpool.

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