Working Paper 2001/2

A COMPARISON OF SYNTHETIC RECONSTRUCTION AND COMBINATORIAL OPTIMISATION APPROACHES TO THE CREATION OF SMALL-AREA MICRODATA

Zengyi Huang and Paul Williamson

October 2001

Department of Geography
University of Liverpool

Abstract

The work reported here offers for the first time a thorough comparison of two established methodologies for the creation of small area synthetic microdata, synthetic reconstruction and combinatorial optimisation. Two computer models, Pop91SR and Pop91CO, have been developed for the reconstruction of ED level populations drawing upon 1991 Census data. The adequacy of their outputs has been assessed at cellular, tabular and overall levels. Consideration has also been given to the impact on outputs of aggregating ED estimates into wards.

Compared with previous synthetic reconstruction models, Pop91SR employs the following new techniques: (a) use of the SAR to examine relationships between variables and determine the ordering of conditional probabilities; (b) a three-level modelling approach to create the conditional distributions, combining data from the SAS, LBS and SAR; and (c) adoption of a modified Monte Carlo sampling procedure. These techniques maximise the use of information and greatly reduce the sampling error, thereby increasing estimation accuracy. The major improvements in Pop91CO are: (a) using a new criterion ($RSSZ_m$) for the selection of household combinations; (b) selection of households from the relevant SAR region, where possible; and (c) a revised set of stopping rules to control the number of iterations and improve the consistency of outputs. Using $RSSZ_m$ as the selection criterion yields significant improvements in the quality of the synthetic data generated.

An assessment of outputs from the two rival approaches, produced using the same small-area constraints, suggests that both can produce synthetic microdata that fit constraining tables extremely well. But further examination reveals that the variability of datasets generated by combinatorial optimisation is considerably less than that for datasets created by synthetic reconstruction, at both ED and ward levels, making combinatorial optimisation the approach of choice for the creation of a single set of synthetic microdata.

Acknowledgements *The work reported in this paper was undertaken as part of an ESRC-funded project on 'The creation of a national set of validated small-area microdata', award no. R000237744.*

A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small Area Microdata

Contents

List of Figures

List of Tables

1. Introduction

Population microdata comprise a list of individuals with associated attributes (e.g. age, sex, marital status, tenure), typically grouped into families and households. The list-based representation of these attributes has significant advantage over array-based representations, including efficient representation, flexible aggregation and data linkage (Birkin and Clark, 1995; Williamson *et al.*, 1998). In Britain the two largest, readily accessible and non-commercial survey microdata sets are the 2% individual and 1% household Samples of Anonymised Records (SAR) from the 1991 Census. Since their release they have been very popular with researchers for area-level and spatial analysis (Dale, 1998). Unfortunately, the types of analyses achievable using the SAR are limited in a number of ways. These limitations include the relatively restricted range of questions asked in the census, the restriction of sample size, the collapsing of response categories, and in particular the restriction of geographical information. In order to protect confidentiality the 1% household SAR contain only a coarse geography (standard statistical region), whilst the 2% individual SAR offers a more detailed district level geography, but at the expense of the loss of a great deal of household level information. Indeed, the lack of spatially detailed information has been recognised as a major limiting factor. King and Bolsdon (1998) pointed out that although local government is a major potential user of the 1% SAR, given its centrality both to local housing policy and to planning policy, this use is greatly restricted by the geography of the SAR.

The need for spatially detailed microdata has been recognised by the Economic and Social Research Council (ESRC). Over the last two decades, ESRC supported research has led to the development of two competing methodologies for producing synthetic small area population microdata, namely 'synthetic reconstruction' and 'combinatorial optimisation'. In this context, 'small area' is taken to mean enumeration district (ED) - the smallest geographical unit in the UK for which census tabulations have been made available. Synthetic reconstruction approach involves the use of Monte Carlo sampling from a series of conditional probabilities, derived from published census tabulations, to create synthetic data. The combinatorial optimisation approach involves the selection of a combination of households from the 1% household SAR that best fit known small area constraints (published census tabulations).

Recently the combinatorial optimisation technique has been examined and assessed by Voas and Williamson (2000a), but to date only partial evaluations of the synthetic reconstruction approach have been made (Williamson, 1995; Duley, 1989). In particular, although initial results have suggested that both approaches have potential, no direct comparison of the results obtained using each approach has been made. In an effort to fill this gap, the aim of this paper is to compare and contrast the two established methodologies (synthetic reconstruction and combinatorial optimisation) for synthetically reconstructing small area microdata, leading to the identification of a favoured methodology as the pre-cursor to the creation of a validated set of national small area population microdata.

The plan of the paper is as follows. Section 2 briefly reviews the two alternative methodologies and highlights their associated problems. Section 3 describes the general approach to evaluating and comparing the two methodologies, in which a set of measures of fit is defined. This is followed by a more detailed description of the two alternative approaches to the recreation of small area microdata. Section 4 is devoted to the development of Pop91SR, a new model based on the synthetic reconstruction approach. A number of technical innovations associated with this approach are reported. Section 5 describes the combinatorial optimisation model, Pop91CO, and various improvements that have been introduced since the writing of Voas and Williamson (2000a). Section 6 presents a thorough evaluation and comparison of the two sets of competing model outputs. The conclusions arising from this comparison are set out in Section 7.

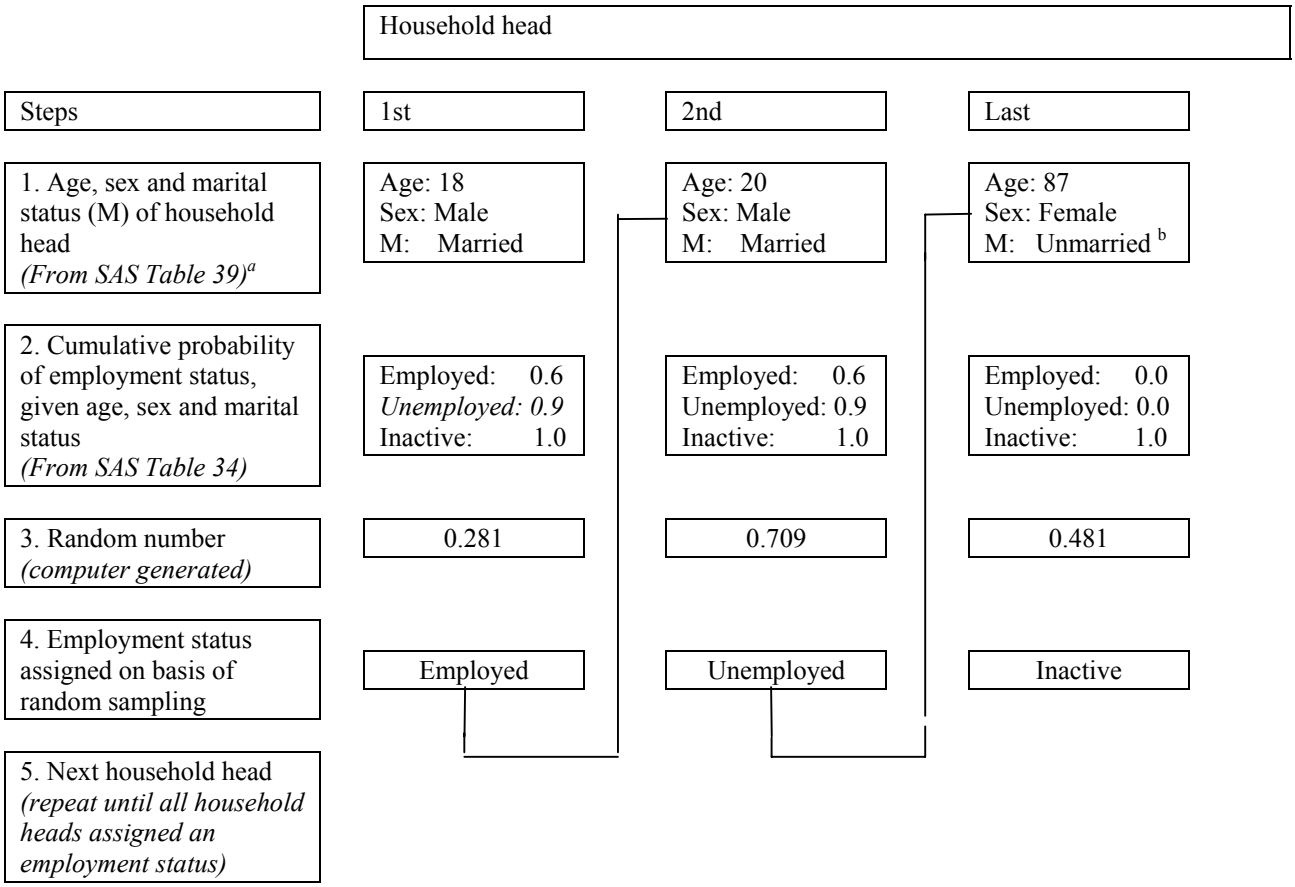
2. Synthetic reconstruction vs. combinatorial optimisation

A number of approaches exist for estimating spatially detailed microdata, including stratified sampling, geodemographic profiling, data fusion, data merging, iterative proportional fitting, reweighting, synthetic reconstruction and their combinations. Given data availability in the United Kingdom, synthetic reconstruction and combinatorial optimisation (a variant of the reweighting approach) have been identified as the two main competing approaches in the creation of small area synthetic population microdata (Williamson *et al*., 1998; Williamson, 2002). Both approaches attempt to create lists of individuals and households whose characteristics are consistent with known aggregate local distributions within the census or other data, but they differ in the means by which they try to achieve this end.

2.1 Synthetic reconstruction

The synthetic reconstruction approach is the most long-standing method for generating synthetic microdata. It normally involves Monte Carlo (random) sampling from a series of conditional probabilities derived from published contingency tabulations. The procedure is usually sequential, and typically begins by creating a set of household heads with age, sex, marital status and spatial location attributes determined by sampling from a known distribution. A next step could be to allocate economic activity to the sample, drawing upon another published tabulation to determine a conditional probability of economic activity given age, sex, marital status, and location of these household heads (see Figure 1). If a head is designated as economically active, a next step might be to estimate his/her occupation. If married, a spouse and, potentially, children could be generated. And so the procedure is carried out for all the variables we wish to include in our synthetic microdata.

The only source of population data at small area scale in the UK and most of other countries is the census, which yields a series of separate, predetermined, aggregate cross-tabulations such as age by sex by marital status for individuals and tenure by ethnic group of head of household for households. It is almost always the case that these tables provide only partial information concerning the conditional chain probabilities to be derived. Suppose that we are interested in the relationships between four variables *a*, *b*, *c*,



[a] Coarse age bands disaggregated into single year of age using other local information
[b] Includes single, widowed and divorced

After Clarke G (1996) 'Microsimulation: an introduction' in G P Clarke [ed.] *Microsimulation for urban and regional policy analysis*, Pion, London, Figure 1.

Figure 1 A simplified synthetic reconstruction procedure

and *d* for a given location and population group, and the interdependencies between these four variables could not be obtained from published census data. But parts of the array are known from the published tables, say $Q_1(a, b, c)$, $Q_2(b, d)$ and $Q_3(a, d)$. In this case,the required conditional probability can be estimated using iterative proportion fitting (IPF), a well-established technique for overcoming data shortfalls of this kind (see Birkin and Clark, 1988; Fienberg, 1970; and Wong, 1992). Specifically, IPF can be used to estimate the full joint distribution, $P(a, b, c, d)$, which fits the constraints Q_1, Q_2 and Q_3. One of the advantages of IPF is that any number of sets of constraints can be embedded within the procedure. If we have created a sample with attributes *a*, *b* and *c*, we can then add variable *d* to the list using the conditional probabilities, $p(d| a, b, c)$, derived from $P(a, b, c, d)$. Hence, through IPF, the data contained in separate tables may be linked together.

In essence, synthetic reconstruction approach tries to reconstruct the original population in such a way that all known constraints (the counts represented in the census tables) are reproduced. A number of models have been developed based on this approach, such as SYNTHESIS (Birkin and Clarke, 1988), UPDATE (Duley, 1989), and OLDCARE (Williamson, 1992, 1996). SYNTHESIS is a model to estimate small area income distribution, OLDCARE is a model of community care services for the elderly, and UPDATE is a dynamic microsimulation model for updating small area populations between censuses. The exact list of variables included in these models depends on the study being undertaken. The majority of variables, such as age, sex, marital status, housing tenure, and socio-economic group, can be estimated using census data. Some variables, such as income in the SYNTHESIS and OLDCARE models, are not available from the census and must be derived from other sources. One main advantage of the synthetic reconstruction approach is that its use of conditional probabilities allows data to be incorporated from the widest possible range of sources.

2.2 Combinatorial optimisation

An alternative way pf estimating small area micropopulations is through the reweighting of an existing large-scale microdata set. The release of the Sample of Anonymised Records (SAR) from the 1991 Census makes it possible to derive list-based estimates of small-area populations by combining information contained in the SAR and the census

small-area tables. The 1% household SAR contains 215,789 households and 541,894 persons resident within those households. Theoretically it is possible to evaluate every possible combination from this SAR and find the set that best fits known small area constraints. But in practice this is almost unachievable owing to computing constraints. For example, the number of possible solutions from the SAR would exceed 10^{690} for an ED with 200 households. Williamson *et al.* (1998) present a combinatorial 'optimisation' approach to offer a way of performing intelligent searching and effectively reducing the number of evaluations. The process is iterative: starting from an initial set of households chosen randomly from the SAR, an assessment is made of the effects of randomly replacing one of the selected households with a fresh household from the SAR. If the replacement improves the fit, the households are swapped. Otherwise the swap is not made. This process is repeated many times, with the aim of gradually improving the fit between the observed data (a set of pre-selected constraining tables) and the selected combination of SAR households. Given the search space, the final combination arrived as is normally the best achievable in a given time, rather than the guaranteed optimal solution. Figure 2 presents a simplified example of this combinatorial 'optimisation' approach.

An initial analysis based on the test of two EDs suggested that the combinatorial optimisation approach produces acceptable population estimates for a suburban ED, where the distribution of constraining tables was close to those of the SAR, but the performance was relatively poor the an inner-city ED, with a distribution markedly different from the national average. Even so, the fit achieved between the estimated population and its constraining tables appeared better than that reported in earlier studies based on the synthetic reconstruction approach (Williamson *et al.*, 1998; Williamson, 1996).

The combinatorial optimisation approach has been further examined by Voas and Williamson (2000a). They developed a 'sequential fitting procedure' to improve the accuracy and consistency of resulting outputs. The most abnormal table for a given area is fitted first (within the target), followed by the next most difficult table, and so on. At each stage changes (household replacements) that favour the fit of later tables at the expense of preceding ones are not allowed. With this sequential fitting procedure they

Step 1: Obtain sample survey microdata and small area constraints

Survey microdata

Household	Characteristics		
	size	adults	children
(a)	2	2	0
(b)	2	1	1
(c)	4	2	2
(d)	1	1	0
(e)	3	2	1

Known small area constraints [Published small area census tabulations]

1. Household size (persons per household)

Household size	Frequency
1	1
2	0
3	0
4	1
5+	0
Total	**2**

2. Age of occupants

Type of person	Frequency
adult	3
child	2

Step 2: Randomly select *two* households from survey sample [(a) & (e)] to act as an initial small-area microdata estimate

Step 3: Tabulate selected households and calculate (absolute) difference from known small-area constraints

Household size	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
1	0	1	1
2	1	0	1
3	1	0	1
4	0	1	1
5+	0	0	0
		Sub-total:	*4*

Age	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
adult	4	3	1
child	1	2	1
		Sub-total:	*2*

Total absolute difference = *4* + *2* = **6**

Step 4: Randomly select one of selected households (a or e). Replace with another household selected at random from the survey sample, provided this leads to a reduced total absolute difference

Households selected: (d) & (e) [Household (a) replaced]

Tabulate selection and calculate (absolute) difference from known constraints

Household size	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
1	1	1	0
2	0	0	0
3	1	0	1
4	0	1	1
5+	0	0	0
		Sub-total:	*2*

Age	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
adult	3	3	0
child	1	2	1
		Sub-total:	*1*

Total absolute difference = *2* + *1* = **3**

Step 5: Repeat step 4 until no further reduction in total absolute difference is possible:

Result: Final selected households: (c) & (d)

Household size	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
1	1	1	0
2	0	0	0
3	0	0	0
4	1	1	0
5+	0	0	0
		Sub-total:	*0*

Age	Estimated Frequency (i)	Observed Frequency (ii)	Absolute difference | (i)-(ii) |
adult	3	3	0
child	2	2	0
		Sub-total:	*0*

Total absolute difference = *0* + *0* = **0**

Figure 2 A simplified combinatorial optimisation process

found it is possible to satisfy a level of minimum acceptable fit for every table used to constrain the selection of households from the SAR.

2.3 Problems of generating small area microdata

Although the previous studies reviewed above suggest that both synthetic reconstruction and combinatorial optimisation are promising approaches, a number of issues are still unresolved. For synthetic reconstruction approach the main problems are:

- *The sampling error*. Synthetic reconstruction of microlevel population data is a Monte Carlo based approach. As a stochastic process Monte Carlo sampling is subject to sampling error. This error is likely to be more significant for small area simulation where the sample sizes are small. The average size of EDs is about 200 households or 450 people. Moreover, our objective is to produce a single synthetic population. Therefore, even if the model's estimates (averaged over many replications) are unbiased the approach may not useful if the variance is too large.

- *The ordering of conditional probabilities*. Synthetic reconstruction is a sequential procedure. A certain amount of error is introduced in each stage, which may be contributed variously by Monte Carlo sampling, modelling assumptions, and data inconsistency. The level of error will increase as we go further along the chain of generation of characteristics. It is thus important to generate new characteristics in an appropriate order so that potential errors are minimised. Because of the lack of an appropriate 'scientific' approach the determination of the ordering relies on the modeller's skills and art (Birkin and Clark, 1995; Clark, 1996).

  Lack of data at ED level. Data are vital for modelling. The models reviewed in Section 2.1 all draw upon the 1981 Census. They suffer from a shortage of census tabulations at ED level and necessarily rely heavily upon data at larger spatial scales (county or national level). The situation has been improved since 1981. The number of available ED level census counts trebled between the 1981 and 1991 Censuses. More detailed data were also made available at ward level, whilst the release of SAR from the 1991 Census offer theoretically limitless flexible tabulations (at least for district level geographies and above). Yet we have not

seen any synthetic reconstruction model making the full use of these information. On the other hand, the increase of available data makes the model building an even more complex task. At each stage some kind of judgement, typically subjective, must be made about the relationships between characteristics. These factors certainly affect the quality of the synthetic data in some way.

The combinatorial optimisation approach provides a novel solution to the problem of generating small area microdata, but there are several areas for further investigation and refinements.

- *Selection criterion*. The existing combinatorial optimisation model uses total absolute error (TAE) as the measure of fit during the iterative fitting process, but the fit of the final synthetic data to the known small area tables is evaluated based on a relative statistic (Z score) (Williamson *et al.*, 1998; Voas and Williamson, 2000a). Whether an alternative iterative fitting criterion would generate improved estimates, and the extra cost (in term of computing time), remains unknown.

- *Stratified sampling from the SAR*. The existing combinatorial optimisation model selects households from the whole SAR. It might be better to limit household selection to households which come from the same region as the small area being synthesised.

- *Table fitting sequence*. The ordering of tables in the sequential fitting procedure is area-specific, and the target level of acceptance is table-specific. These constraints make it difficult to apply the sequential fitting procedure in generating large area microdata.

For both approaches appropriate measures of fit between the synthetic population and the known constraints are debatable (see Voas and Williamson, 2001a). In addition, for models that work from the bottom-up (starting at ED level), the assessment of fit is complicated by among other factors: (1) the tables of observed and expected results are often sparse, i.e. many counts are at or close to zero, and (2) an observed count may not be the actual count. The latter is the result of data blurring applied to the released ED and

ward census tabulations for confidentiality reasons. With the exception of a few basic counts of total households and total population, non-zero counts in the census Small Area Statistics (SAS) have been modified by the addition of +1, 0 or –1 in quasi-random patterns. This leads to discrepancies in census counts between tables and between ED and ward totals.

The inconsistency between the constraining tables could also cause problems in modelling. The effect for the synthetic reconstruction approach may be significant because convergence will not occur during IPF where there is a mismatch in the table totals or subtotals. For the combinatorial optimisation approach the impact of data blurring may mean there is no possible combination of households that would match every constraining table perfectly.

In summary, synthetic reconstruction is a well-established approach for the creation of synthetic small-area micriodata. The increase of small area data and the release of census microdata at coarse spatial scale offer great potential to build better models. Combinatorial optimisation is a promising alternative to the creation of synthetic microdata, although further refinements could be envisaged. In the light of above, the objectives of this paper are twofold: first, we attempt to explore techniques to tackle some of the problems described above in order to improve the resulting outputs of both approaches; and second, we evaluate and compare the two competing methodologies leading to the identification of a favoured methodology for generating small area microdata.

3. Approach to evaluation

In order to evaluate and compare the two main competing approaches, two models (Pop91SR and Pop91CO) have been developed for the reconstruction of ED level population microdata, drawing upon 1991 Census data. Pop91SR is a new programme suite based on the synthetic reconstruction approach. Attention has been paid to establishing a procedure for determining the ordering of the conditional probabilities, through use of the SAR, increasing the accuracy of the estimated conditional distributions and reducing sampling error. Pop91CO is the latest version of the alternative combinatorial optimisation programme suite. Effort has been focused on determining the best criterion for household selection and improving the quality and consistency of the synthetic dataset.

3.1 Assessing performance

The performances of each approach will be assessed with respect to their effectiveness and efficiency in generating synthetic datasets. The main element in evaluating performance is the ability to produce accurate and reliable spatially-detailed microdata (effectiveness). Synthetic microdata will not be identical to the actual records of households and individuals from the area in question, and any evaluation should identify the nature and extent of these discrepancies. Because a full set of raw census data is not available to us, we are obliged to evaluate the reconstructed populations by comparing a number of aggregated tables derived from the synthetic dataset with published census small-area statistics. The principle is that if we examine enough cross-sections of multi-dimensional space, we can assess the overall resemblance between the datasets themselves. In what follows we discuss the major factors that should be considered in accessing the reliability of results and our adopted approaches to evaluating the alternative synthetic datasets.

- *The stochastic nature of the models.* Both synthetic reconstruction and combinatorial optimisation are stochastic processes. Variations in the sample seed value will alter the random number string and so alter assignments (or household selection) and the estimated counts. Knudsen and Fotheringham (1986) suggest that if the object of the analysis is to assess the accuracy of a model in replicating a single dataset,

significance tests need to be undertaken. In this study the assessment of model results are based on 100 replications. Each run starts with a different initial sample seed. This allows us to test model performance on each synthetic dataset individually, and on all datasets in aggregate, thereby giving both the fit of mean and the mean fit (i.e. the bias and the average error).

- *Measures of fit.* The discrepancies between estimated and observed tabular data can be looked at in different ways (Voas and Williamson, 2001a). Some measures emphasis on absolute differences, others assess relative distributions. Consequently, a synthetic dataset may show a good fit by one measure but not by another. The fit may be good overall, but poor at a particular point. Information on bias and variability at the cellular level will, therefore, be as important as at tabular level. Consequently, in order to undertake a thorough evaluation of model outputs, we have designed a set of test statistics that measure the fit of the synthetic microdata to local constraints at cellular, tabular and overall levels (detailed in Section 3.2).

- *Geographical scale.* Intuitively, estimates at ED level could be subject to a considerable degree of error, particularly if they are based on a single run. Hence, we may wish to examine whether or not error can be balanced across EDs. The fit might relatively poor at ED level but much better at higher geographical levels. Conversely, fit at ED level does not guarantee a perfect fit at a coarser spatial scale. Previously, this element of fit has received little consideration. In particular, neither of the sets of synthetic 1981 British population microdata described in section 2 have been subject to multi-geographical level evaluation. In section 6 we test the fit of synthetic datasets at both ED and ward levels using the same set of test statistics. Attention is also paid to establishing the degree of variability associated with a range of estimated count data at both levels.

- *Comparability of outputs*. Initially at least, the same set of ED level constraints are used by both the synthetic reconstruction and combinatorial optimisation approaches. This allows us to directly compare and evaluate the fit of the two sets of synthetic data using the same set of ED level constraining tables. For both combinatorial optimisation and synthetic reconstruction datasets based on ED data only, the extent

to which known (but unused) ward level interactions have been captured can also be assessed. However, the ideal synthetic reconstruction model uses these ward level data, because they fill known gaps at ED level. For instance, data linking household heads' demographic characteristics with economic position or housing tenure are only available down to ward level. Consequently, we have created a second synthetic reconstruction based dataset using ward level tables as additional constraints. This second dataset allows the impact on ED level fit of adding ward level constraints to be evaluated for the synthetic reconstruction approach.

- *The selection of test areas*. Previous studies (Duley, 1989; Williamson *et al.*, 1998; Voas and Williamson, 2000a) have shown that the quality of synthetic microdata produced by either synthetic reconstruction or combinatorial optimisation varies with location. Usually the fit is good when the distribution of local constraining tables is similar to the overall population's, whilst the performance is less good if the tables are atypical relative to the national distribution. In particular the greatest population estimation problems have been encountered when dealing with inner-city and 'student' areas. Accordingly, test areas should cover different types of EDs and include both 'normal' (close to norm) and 'abnormal' (far from the norm) areas. We have selected two wards, the Cookridge and University wards of Leeds, as test areas. The former may be described as a typical suburban area and the latter as an inner-city and 'student' area. As a test of whether these wards meet our criteria, we can compare the distributions of the EDs in the two wards with that of all the EDs in England and Wales according to their distance from the norm, using a standardised measure based on 54 census variables (described in Voas and Williamson, 2000b). The results are shown in Figure 3. Clearly, the test areas comprise EDs of different types, which are well dispersed over the range of the national distribution. The distribution of the EDs in the Cookridge ward is very similar to the national one. In contrast, most of the EDs in the University ward are far from the norm. Nearly half of them (47%) lie outside the 90th percentile of the national distribution; 23% are extremely atypical (outside 99th percentile); and two are among the top four furthest EDs from the norm in England and Wales (Voas and Williamson, 2001b: Table II). This demonstrates that the two wards selected offer a reasonable test: one is quite 'normal' and the other extremely 'abnormal'.

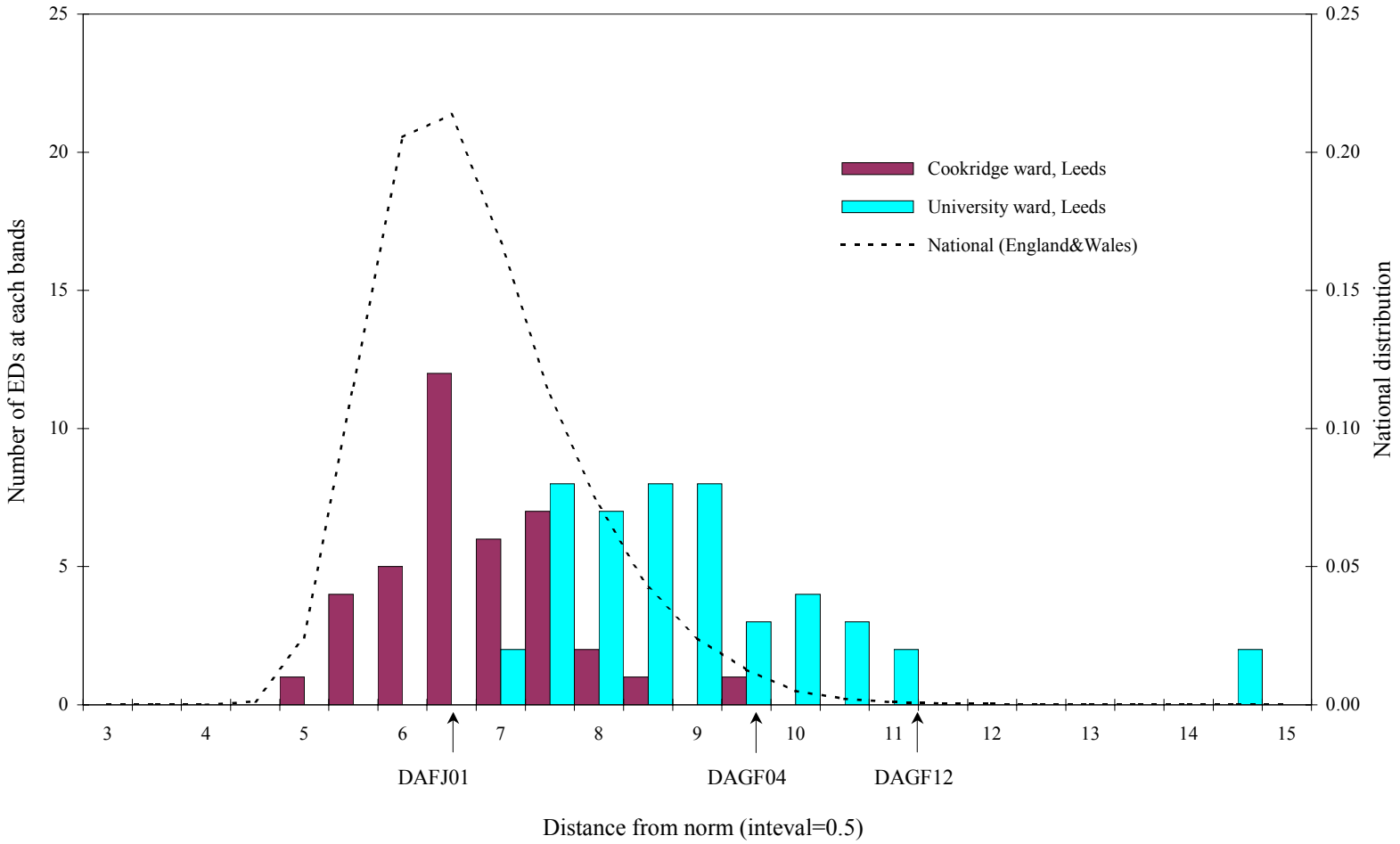


Figure 3 The distance of test EDs from the national norm

Apart from examining the reliability of outputs, it is necessary to assess the efficiency of each approach. Bearing in mind that our objective is to identify the best approach for creating population microdata for large areas, from metropolitan district level to a whole nation, the resource costs of generating such data should be assessed. We consider the two main inputs: man-hours for developing the model and computing time for running the model. They are likely to be very different for the two approaches. The development of a small area population reconstruction model typically takes a considerable period of time (person months), but takes considerably less time to run than the combinatorial optimisation model. These factors will be assessed in conjunction with the model's reliability in section 6.2. It is also desirable to estimate the extra cost of adding more variables and constraints, or altering the exist set of constraints. Due to time limitations, we can only select a basic set of attributes in our datasets, which we believe is sufficient for the purpose of comparison. Others may wish to include more variables or change some of them in a final synthetic dataset, and the flexibility of adding or altering variables and constraints becomes an important aspect in the assessment of a model's efficiency.

3.2 Testing statistics

Numerous statistics have been used to assess model goodness-of-fit, but the choice remains difficult. Knudsen and Fotheringham (1986) classified these statistics into three types: information-based statistics, general distance statistics, and traditional statistics. Information-based statistics have their origin in the information gain statistic; these include the phi statistic, the psi statistic, the absolute value formulation of the psi statistic, and the absolute entropy difference. General distance statistics simply measure the differences between observed and estimated counts. The differences are either squared or made absolute to avoid summing positive and negative values. A representative of these statistics is the standardised root mean square error (SRMSE). Traditional statistics include R^2 and the chi-square statistic. According to Knudsen and Fotheringham's study, for analysing the performance of two or more models in replication of the same data set, or for comparing a single model in different systems, the most accurate statistics appear to be SRMSE, the absolute value formulation of the psi statistic and the phi statistic. The chi-square statistic is particularly poor for these purposes.

There are, however, several problems with employing the statistics suggested by Knudsen and Fotheringham in our analysis. SRMSE should only be used when the total of the estimated table equals that of the observed table (Knudsen and Fotheringham, 1986:132). Because the synthetic dataset may not contain the same number of individuals as the actual population, and because census data are modified before release to protect confidentiality, totals in the tables being compared will not necessarily match. The use of information-based statistics such as psi and phi statistics are also problematic because they remain undefined when one or more of the observed counts equal zero. Small area tabulations often contain many values at or close to zero. Finally, the use of traditional tests based on chi-square are subject to similar problems. They are designed to test the goodness-of-fit of an entire table rather than individual cells, are also subject empty cell problems, and perform poorly on relatively large sparse tables.

In the light of these difficulties, a number of authors have used the 'Z-statistic' or its modified version to test the fit of individual counts, cell by cell (Birkin and Clarke, 1988; Duley, 1989; Williamson, 1992; Williamson *et al.*, 1998). Recently Voas and Williamson (2001a) offered an in-depth appraisal of the Z-statistic and its variants, along with other measures. They put particular emphasis on the suitability of measures for the evaluation of synthetic microdata. The conclusions of their study may be summarised as follows:

(1) The most straightforward test statistic is total absolute error (TAE), which is calculated simply as the sum of the absolute differences between estimated and observed counts. Although relatively crude as a measure of fit, it is easy both to calculate and to understand. It can be used to compare rival models against the same table, but not performance across different tables. The standardised absolute error (SAE), which is TAE divided by the total expected count for the table, is marginally preferable on the grounds that it may be valuable for quick and easy comparisons across tables.

(2) The forms of the phi and psi statistics are shown to be closely approximated by the simple measure of absolute error, SAE. Since SAE is simple and readily understood, there seems to be little benefit in using phi or psi as an alternative.

(3) The preferred measure is a normal Z score for each table cell. The Z score is based on the difference between the relative size of that category in the synthetic and actual populations, although an adjustment is made to the formula when dealing with zero counts. The Z score is preferred because it has known statistical properties, wide acceptance as a valid measure of fit, and can assess not only cellular fit but also, when aggregated, tabular fit (see (5)).

(4) The modified Z score (Z_m) proposed by Williamson *et al.* (1998, Appendix) is recommended for use during the iterative fitting process (combinatorial optimisation approach only), since the synthetic totals may not be identical to the actual total. In cases where observed and synthetic totals are the same, $Z_m = Z$. As the final synthetic and target totals will be highly similar, an unmodified Z score should be used for evaluating end results.

(5) The Z statistics for individual counts, when squared and summed, provides a measure of fit for the entire table. The sum of squared Z scores (which we will label SSZ) has a χ^2 distribution with degrees of freedom equal to the number of table cells (Voas and Williamson, 2001a). If a table's SSZ exceeds the table-specific 5% χ^2 critical value, then the dataset is deemed not to fit (an application of this statistic can be found in Voas and Williamson, 2000a).

Based on the results of this study and our considerations as described previously, we have designed a set of test statistics for comparing the two sets of synthetic data (see Table 1). At the detailed level, for every count we calculate the mean estimate, the estimated 95% confidence interval (i.e. the range within which 95% of the synthetic values fall over 100 replications), the normal Z score and the Z score of the mean estimate. If a cell produces a synthetic count with normal Z score exceeding the 5% critical value (i.e. $|Z|>1.96$), then the synthetic data is deemed not to fit that cell. Such a cell is called a 'non-fitting cell' (NFC). However, as already noted, ED level census data (SAS tables) have been modified by randomly adding +1, 0 or –1, the effect of which is most severe when the value of the count is small. Unfortunately, it is difficult to separate the error contributed by data blurring from that caused by the estimation process itself. One simple way of attempting to assess the possible impact of data blurring is to assume that the actual count

could be one greater or one less. Consequently, for every count we also calculate two other Z-scores by adding +1 and –1 to the SAS count, which are denoted by Z_{+1} and Z_{-1} respectively. If a cell's Z score exceeds the critical value but either Z_{+1} or Z_{-1} value does not, the discrepancy could theoretically be caused, at least in part, by data blurring. In contrast, where a cell produces a synthetic count with all Z, Z_{+1} and Z_{-1} scores exceeding the critical value, then the fit is undeniably poor and the cell is designated as a 'poorly-fitting cell' (PFC).

The measures we have adopted for tabular fit are aggregations of statistics calculated for the individual cells. We use three types of test statistics to assess the magnitude of the discrepancies between the estimated and observed tables: the number of non-fitting and poorly-fitting cells per table; TAE (total absolute error); and SSZ (sum of squared Z-scores). The numbers of NFC and PFC per table are simply the sums of the cellular test results. TAE provides a quick and easy measure of absolute differences between observed and synthetic table counts. The more complex SSZ assesses proportional differences and provides a more robust appraisal of tabular fit. If a table's SSZ exceeds the table-specific 5% χ^2 critical value, it is deemed to be a 'non-fitting table' (NFT). Note that in this case no allowance is made for the possible adverse impact of data blurring, as the overall impact of data blurring on a given table should be broadly neutral. In addition, measures of tabular fit have been developed to assess fit over 100 replications. The NFT-rate simply records the number of times out of 100 replications that a table is designated as a non-fitting table on the basis of its SSZ. In contrast, the SSZ of mean identifies the fit of a table's 100-run mean, again on the basis of SSZ.

When assessing the fit of a synthetic dataset to a set of constraining tables, it is convenient to have some measure of overall fit. Ideally, such a measure of overall fit would be a simple aggregation of tabular test results. One obvious candidate is SSZ. A difficulty encountered is that at tabular level the magnitude of SSZ depends on the number of table cells, as well as the degree of error. The bigger a table, the larger the value of SSZ is likely to be. The solution adopted is to divide a table's SSZ by the table-specific 5% χ^2 critical value. We call this new statistic the relative sum of squared Z scores, or RSSZ. RSSZ appears to be more informative than SSZ. First, as the table-specific critical values already take into account the number of table cells, it provides a

Table 1 Test statistics for evaluating the fit of synthetic microdata

Cellular level	Tabular level	General level
Mean synthetic	TAE	Overall TAE
95% confidence interval	SSZ	Overall RSSZ
Z score of mean	SSZ of mean	Overall RSSZ of mean
-	NFT rate	Number of NFT
Z score	Number of NFC	Overall number of NFC
Z_{+1} and Z_{-1} scores *	Number of PFC *	Overall number of PFC *

* Only used for comparing the synthetic data with SAS tables.

relative measure that can be used to assess the performance across different tables. Second, the RSSZ statistics for individual tables, when aggregated, provides a measure for overall fit for a set of tables that treats fit to each table with equal importance. Third, the value of RSSZ is simple to interpret; if it is less than one then the data fit the table. Other measures to assess and compare the overall fit of the alternative synthetic datasets are shown in column 3 of Table 1, all based on summing various measures of tabular fit already discussed above. The first three statistics (TAE, RSSZ, and RSSZ of mean) are mainly used for the comparison of the two synthetic datasets, while the last three measures (numbers of NFT, NFC, and PFC) are better used to identify key sources of error, although they can be used for overall comparison purposes as well.

At ward level comparisons of synthetic estimates with census counts are considerably affected by the census data blurring process. Due to data blurring, the ward sum of ED level counts for a given table cell are only guaranteed to fall within $\pm n$ of the published ward level count, where n = number of EDs in the ward. As ED level counts are used as the basis for creating synthetic microdata, comparing aggregated ward level synthetic microdata to published ward level counts could be very misleading. To overcome this problem comparisons are made instead with SAS table counts aggregated to ward level. However, if a cross-tabulation we wish to examine is available only at ward level, then necessarily that table will be used. Given uncertainties over the impact at ward level of the data inconsistencies between tables caused by data blurring, the concept of poorly-fitting cells (PFC) is not used at ward level.

4. The synthetic reconstruction model (Pop91SR)

In this section we describe the development of Pop91SR, a synthetic reconstruction model for recreating small area population microdata with the use of 1991 census data. The population group considered by Pop91SR is residents in households (household residents). Residents in communal establishments are excluded, as is the case for the combinatorial optimisation approach reviewed in section 5. In this section, we first consider the available data and their suitability for synthetic reconstruction. This is followed by a description of the methodology for synthetic population reconstruction. Several technical innovations designed to increase the accuracy of estimation are introduced. Finally, the construction process of Pop91SR is presented in a step-by-step description of both inputs and outputs.

4.1 Data for the construction of Pop91SR

Three types of data from the 1991 Census have been used for the construction of Pop91SR: the Small Area Statistics (SAS), the Local Base Statistics (LBS) and the Samples of Anonymised Records (SAR). The only available data at ED level comes from the Small Area Statistics (SAS), which comprise a set of 86 tables providing aggregate data. More detailed information are available at coarser spatial scales. One of the major innovations introduced with the 1991 Census is the expansion of the local statistics output into two separate but interrelated sets (Dale and Marsh, 1993:205). The lower tier is the SAS and the upper tier is the LBS. The LBS consist of 99 tables available down to ward level. The SAS are in fact an abbreviated version of the LBS. The SAS comprise about 9,000 statistical counts for each area, while the LBS comprise about 20,000 counts. Not only are some of the LBS tables omitted in the SAS but also, in most cases, the level of detail in an LBS table is reduced when producing a corresponding SAS table. Both LBS and SAS are available in machine-readable form.

The quality and suitability of the SAS and LBS for small area population reconstruction can be examined from the following dimensions: the spatial scale, the number of variables cross-tabulated, the number of classes within a variable, the effect of data modification, and the sample size. The SAS tables have the most spatial detail and are vital for providing small-area constraints to the synthetic reconstruction process. The

LBS tables are useful for two main reasons: first, they can provide extra constraints on relationships between variables that are not available at ED level, and second, they can be used to disaggregate coarser ED level data to finer classifications.

Census tables vary greatly in size. For instance, the number of cells in a SAS table range from a dozen to nearly 200. This is because of the differences in the number of variables involved and the detail of their categorisation. From the viewpoint of small area population reconstruction, larger tables are more useful because they provide more detailed local information. However, very large tables should be used with care, particularly at ED level, because the larger a table, the smaller the individual cell counts are likely to be and the greater the possible impact of data blurring. For example, adding or subtracting 1 from 2 has a far greater proportionate effect than it does on values of 20 or 200. But by far the major factor affecting the suitability of a SAS/LBS table is sample size. Most of the SAS and LBS tables report on variables which are 100% coded; but a small set reports on variables which are coded only for 10% of census returns. The 10% SAS have only been released at ED level to provide users with a primary building block to estimate 100% population of much larger areas. They are not suitable for produce 100% counts of small area because of their large sampling error.

The SAR are anonymised microdata extracted from the 1991 census. They can serve two main purposes. First, the SAR can be used to examine the relationship between variables, helping to determine the ordering of chain probabilities. Second, the SAR can be used to strengthen weak links in the reconstruction process. The SAR can be aggregated in any way to produce joint distributions that are not available in the SAS/LBS. For synthetic reconstruction use of both 2% individual and 1% household SARs can be appropriate, but in this study we use only the 1% household SAR throughout. The SAS, LBS and SAR together provide a rich data source for the reconstruction of small area microdata.

4.2 Methodologies

4.2.1 Iterative proportional fitting

The synthetic reconstruction procedure to generate population microdata from a variety of aggregate data is underpinned by the method of IPF. The theoretical aspects of IPF have

been investigated thoroughly (e.g., Fienberg, 1970; Bishop *et al.*, 1975), and the utility and reliability of the procedure in geographical research and population modelling have been demonstrated (e.g., Wong, 1992; Norman, 1999; Birkin and Clark, 1988; Duley, 1989; Williamson, 1992). In a simple case, the IPF procedure can be used to ensure that a two-dimensional table of data is adjusted so that its row and column sums equal to predefined values. Let $P^k(i,j)$ be the matrix element in row i, column j, and iteration k. $Q(i)$ and $Q(j)$ are the predefined row sums and column sums. Starting from an initial matrix $P^0(i,j)$, the new cell values are estimated iteratively by the following set of equations:

$$P^{k+1}(i,j) = \frac{P^k(i,j)}{\sum_j P^k(i,j)} Q(i) \tag{1}$$

$$P^{k+2}(i,j) = \frac{P^{k+1}(i,j)}{\sum_i P^{k+1}(i,j)} Q(j) \tag{2}$$

The iterative estimation process will stop when convergence to the desired accuracy is attained. A satisfactory stopping rule is to decide on a quantity δ (e.g., $\delta = 0.001$) and stop when a complete cycle does not cause an cell to change by more than this amount, that is, when

$$| P^{k+2}(i,j) - P^{k+1}(i,j) | < \delta \tag{3}$$

for all i and j.

In the context of population reconstruction, $Q(i)$ and $Q(j)$ could be the total population in the ith and jth categories of any two census variables, and $P(i,j)$ will be the estimated values in the cross-classified categories defined by the two variables. The same principle can be applied to estimate an n-dimensional array when only partial distributions or marginal totals are available. Examples of the IPF procedure for three variables have been given by Bishop, *et al.* (1975:84). For more detailed discussion of using IPF to estimate conditional probabilities see Birkin and Clark (1988).

During the construction of Pop91SR, the IPF procedure has been mainly used in two situations: (1) inflating 10%-based tables and (2) augmenting joint count or probability distributions.

4.2.2 Inflating 10%-based tables

Data on some variables we wish to include in our sample may be only available in the 10% form. Although 10% SAS are not modified, unreliable results would be produced if we simply multiplied the 10% SAS counts by 10; they are only released at ED level to provide a basis for flexible area aggregations (Dale and Marsh, 1993:230). Alternatively, we can use the corresponding ward-level 10% table to estimate ED level table counts. For example, during the population estimation process we have to use a 10%-based table, SAS table 86, which gives the break down of the socio-economic group of household heads by tenure. The socio-economic group is a 10% coded variable, which is divided into nineteen categories including economically inactive. To estimate the 100% counts of this table, a two-stage estimation is adopted. The first stage is to estimate the 100% counts of the table at ward level. Let $P_w^0(s,t)$ be the matrix element in row s (socio-economic group of household heads) and column t (tenure), given by LBS table 86. $Q_w^0(s)$ and $Q_w^0(t)$ are the row sums and column sums. The marginal totals $Q_w^0(t)$ are not reliable because they are 10% sample and do not include imputed households. The distribution of tenure, denoted by $Q_w(t)$, is available from other 100%-based tables (e.g., LBS table 42). No better data are available for the socioeconomic group of household heads, so $Q_w^0(s)$ is weighted in such a way that the total is equal to that of $Q_w(t)$. Let $Q_w(s)$ be the weighted distribution of the socioeconomic group of household heads. Then the IPF procedure is used to estimate the ward level array $P_w(s,t)$ constrained by $Q_w(s)$ and $Q_w(t)$ given an initial input of $P_w^0(s,t)$.

In stage two, we estimate the ED level cross-distribution of these two variables, $P_e(s,t)$. The ED level tenure distribution $Q_e(t)$ can be obtained from 100%-based table (e.g., SAS table 42. Although the marginal totals of the socioeconomic group of household heads presented in SAS table 86 is unreliable, when aggregated they may provide some vital local constraints. From this table we can estimate the proportions of household heads who are economically active and who are economically inactive at ED level, a constraint that is not available in other SAS tables. Let $Q_e(s')$ be the ED level distribution of the

socioeconomic group of household heads, divided by two categories: economically active and economically inactive. Similarly, the IPF procedure is used to estimate the ED level array $P_e(s,t)$ constrained by $Q_e(s')$ and $Q_e(t)$ given ward level array $P_w(s,t)$ as the initial estimates.

4.2.3 Augmenting cross-classifications

The main function of the IPF procedure in synthetic population reconstruction is to estimate augmented joint distributions or conditional probabilities for the variables of interest. A three-level estimation procedure is adopted to create an ED level joint distribution, which incorporates data from the SAS, LBS and SAR.

1. Estimate the national level joint distribution for the all variables concerned derived directly from the SAR;
2. Estimate the ward-level joint distribution for the variables concerned given the national level distributions and ward level constraints (the LBS tables), using the IPF procedure;
3. Estimate the ED-level joint distribution for the variables concerned given the ward-level distributions and ED level constraints (the SAS tables), using IPF.

This process is illustrated in Figure 4.

An example is useful at this point to clarify the process. Suppose we have created a synthetic population of household heads for a given ED with the characteristics of age (a_1), sex (g) and marital status (m_1). The subscript number 1 refers to the target set of variable categories (e.g., a_1 may indicate age breakdown is by single year); larger numbers represent coarser disaggregations. This sample is constrained by the known ward level distribution $Q_w(a_2, g, m_2)$ and ED level distribution $Q_e(a_3, g, m_3)$. We now wish to add another attribute, say x, to the synthetic data. Our target is to estimate the conditional probabilities $p(x_1|a_3, g, m_3)$, i.e., the probability distributions of attribute x_1 given the household heads' coarse age group (a_3), sex (g) and coarse marital status (m_3).

The problem can be viewed as needing to estimate an ED level joint distribution $P_e(a_3, g, m_3, x_1)$. A cross-tabulation of these four variables can be derived from the SAR, which

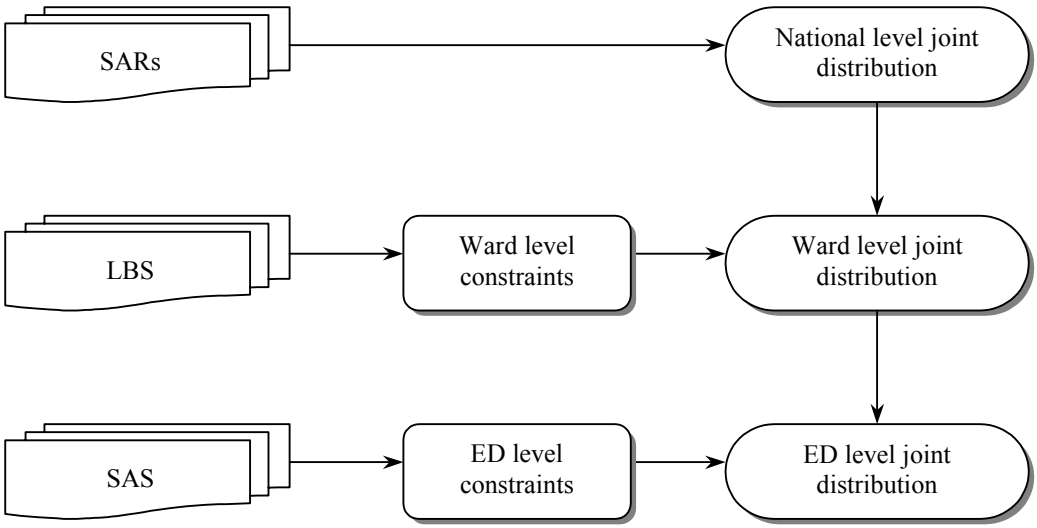


Figure 4 A three-level estimation procedure

acts as a national level distribution. So we have an array $P_n(a_3, g, m_3, x_1)$ at national level. From the LBS we may have a table, $Q_w(a_3, g, x_1)$, which links two of the existing variables with the new one. This table and the previous one, $Q_w(a_3, g, m_3)$, aggregated from $Q_w(a_2, g, m_2)$, are the ward level constraints. Note that a variable may be grouped in different ways in different tables. Using the IPF procedure we can adjust $P_n(a_3, g, m_3, x_1)$ to fit $Q_w(a_3, g, x_1)$ and $Q_w(a_3, g, m_3)$, resulting in an estimated ward level distribution $P_w(a_3, g, m_3, x_1)$.

At ED level, we may find only one table from the SAS linking one of the existing attributes, say age, with variable x, and both variables are in coarser groups, $Q_e(a_3, x_2)$. In a similar way, we can weight the ward level distribution $P_w(a_3, g, m_3, x_1)$ to fit the ED level constraints $Q_e(a_3, x_2)$ and $Q_e(a_3, g, m_3)$, and obtain an estimated ED level distribution $P_e(a_3, g, m_3, x_1)$.

Here, IPF acts as a weighting system whereby the elements of an array of a higher geographical level are adjusted iteratively (scaled down) to fit known constraints of the lower geographical level. The resulting array will retain the interaction pattern of the higher level one, where unknown at lower levels. In general, an increase in the amount of information included via the constraints will improve the accuracy of the estimation. Employing the three-level estimation procedure, it is easier to identify all relevant information from the census and include them as constraints. It also reduces the complexity of dealing with two-level constraints at the same time. Another advantage of this approach is that we can generate a ward-level dataset at the same time, though this is beyond the scope of this study.

Birkin and Clark (1988) summarised the outcome of the IPF procedure as follows:

- All known information is retained, and may be generated anew via reaggregation.
- Although no new information is actually generated, maximum likelihood estimates are provided for missing cell probabilities.
- Any model incorporating partial information may be treated in this way. In practice, the maximum possible information should be included through the constraints.

They also noted that no errors are introduced by the IPF process (i.e. we can estimate a complete set of joint probabilities which is completely consistent with all known constraints) (Birkin and Clark, 1995, p373). But this is subject to the comparability of the constraints. Any overlap between the constraints must be consistent. For example, if two constraints $Q(a, b)$ and $Q(b, c)$ are to be fitted, the common vector b must be identical. Otherwise, the convergence of IPF will not occur (see Bishop et al, 1975, 101).

A key challenge to the use of IPF in small area population reconstruction is that inconsistency exists between constraining tables due to data blurring. It is necessary to adjust the constraining tables so that there are no inconsistencies between them before using the IPF procedure. Fortunately a few tables containing basic counts such as the numbers of households and resident in households within an ED are unmodified, abd provide a basis for adjustment. Each constraining table is subject to one of the following types of adjustment.

(1) *Adjust table to fit known total*. For instance, SAS table 39 (S39) gives the breakdown of the number of household heads by age, sex and marital status, which is the first SAS table used to generate a sample of household heads. The table total is checked against the known number of households for that ED. If they are not identical, the difference is randomly added to or subtracted from table cells according to the relative size of cell. The adjusted S39 total fits the known total.

(2) *Adjust table subtotals*. Sometimes one of the table variables has already appeared in a previously adjusted table (e.g., age of household heads). In this case, the new table elements are adjusted to fit 'known' subtotals. Similarly a three-dimensional table can be adjusted to fit the 'known' cross-tabulation of two dimensions.

(3) *Reconcile variant cell counts*. A special case is the scenario in which two tables contain the same variables, but the population group of one table is a subset of the other. For example, S35 and S39 give the cross-tabulation of age, sex and marital status for household residents and household heads respectively. There are more age groups in S35 compared with S39. If we aggregate S35 to the format of S39, it is quite often found that in one or two cells the number of household heads is larger than that of the residents, due to data blurring. It is not possible to judge which

count is correct, so in this case we assume the S39 count is correct and adjust the S35 count to match.

In reality these various adjustments are unlikely to bring about much perturbation into the SAS. Comparing the total of S39 with the actual number of household heads for every ED within the two test wards, it is found that the maximum net error is two, and for 92% of EDs the net error is either one or zero. Therefore, for the majority of EDs S39 is unchanged or just one cell is altered by +1 or -1. Type (2) or (3) adjustment may alter more cells, but the difference between a SAS table and the adjusted one is again likely to be trivial. Although an adjusted table may be not more accurate than the SAS table, the adjustment process guarantees that at least the total and subtotals of constraining tables are consistent.

Another issue related to the use of IPF is the complexity of the process. In the past when anonymised census microdata were not available, a new attribute was normally made directly dependent upon only two or three existing attributes. Using the SAR it is possible to estimate any joint distribution of census variables desired, with large multivariate joint distributions from the SAR coded to district or above constrained to local conditions using ward and ED SAS tables. Consequently, as more variables are added to the synthetic microdata, many constraining tables with different categorisation schemes become involved, and the IPF procedure becomes extremely complex. A technique has been developed to reduce this complexity, which will be discussed in Section 4.3.

4.2.4 Random sampling: a modified procedure

Having derived the necessary conditional probabilities, synthetic microdata is created through the use of random (Monte Carlo) sampling. This method involves the generation of a string of pseudo-random numbers to assign attributes on the basis of sampling from the relevant conditional probabilities (as shown in Figure 1). At the first step of the population reconstruction process, the Monte Carlo method is used to *disaggregate* data. After that it is used to *augment* data.

A fundamental characteristic of the bottom-up approach is that the separate, aggregate data at lower geographical level are linked and disaggregated with detailed data at higher geographical level. To begin with, for a given ED we create a sample of household heads with the characteristics of coarse age (a_3), sex (g) and coarse marital status (m_3), which can be obtained directly from the adjusted S39. This table is disaggregated using ward level tabulation $Q_w(a_2, g, m_2)$ from L39 and national level tabulation $Q_n(a_1, g, m_1)$ from the SAR. So we have the conditional probability distribution $p(a_1, m_1|a_3, g, m_3)$. Monte Carlo sampling is used to disaggregate our sample into the target set of categories. Next we create the conditional probabilities $p(x_1|a_3, g, m_3)$ for the new variable x. Monte Carlo sampling is then used to augment our sample with variable x. The same process is used for all attributes.

As a stochastic process Monte Carlo sampling is subject to sampling error. This error is likely to be more significant for small areas where the sample sizes are small. Huang and Williamson (2001) presented a modified sampling procedure designed to reduce this sampling error. The procedure can be summarised as follows:

(1) Partition the synthetic population into groups that match the cells in the conditional probability to be used for adding/disaggregating a variable.

(2) Calculate the target distributions for each group by multiplying the conditional probability by the group total.

(3) Separate every count of the target distribution into integer and fraction parts. Use Monte Carlo sampling to turn the fraction distribution into an integer one.

(4) For each group, assign each member in the group to a category of new variable (or finer categories) according to the target integer distribution.

For example, we know the probability distributions of variable x (x_1), given the household heads' coarse age (a_3), sex (g) and coarse marital status (m_3). So, in step one our sample is divided into groups according to household heads' age (a_3), sex (g) and marital status (m_3). In step two the target distribution is calculated. Suppose x_1 contains five categories and the probability distribution of this variable for a given group is {0.12, 0.25, 0.52,

0.04, 0.07}. If the group size is 20, then the target distribution of variable *x* for the group is {2.4, 5.0, 10.4, 0.8, 1.4}. In step three, the fraction part of this distribution is separated, which is {0.4, 0, 0.4, 0.8, 0.4} with the sum of 2. Monte Carlo sampling is used to turn this distribution into an integer one, say {1, 0, 0, 1, 0}, i.e., assigning one person in the first cell and one in the forth cell. We now have a target integer distribution {3, 5, 10, 1, 1} for this group. The final step is randomly assigning these 20 people to a category of variable *x* to match this distribution.

The advantage of modified over conventional Monte Carlo sampling is that only the fractional part of the target distribution is subject to random sampling. As a result, modified sampling can produces more accurate estimates. The degree of improvement generated by this procedure depends upon the relative size of the fraction (or integer) part of a target distribution. The smaller the fraction part, the less random variability remains and, therefore, the greater the degree of improvement in comparison to conventional Monte Carlo sampling. Where all the target values are integers no error will occur from the modified sampling (in term of matching the two distributions). Where all the target values are less than one, the modified sampling procedure becomes the same as non-modified sampling.

4.3 The Pop91SR reconstruction process

As discussed in Section 2, the synthetic reconstruction of population is a step-by-step process. The value of each individual or household characteristic is estimated by random sampling from a probability conditional upon one or more previously generated attributes. Pop91SR starts by generating a set synthetic heads of household with the characteristics of age, sex, marital status and location (enumeration district). This is based partly upon the assumption that location, age, sex, and marital status of the head of household are the most fundamental characteristics of household structure, but also on practical considerations of what data are available, since these features are all cross-classified at ED level (S39). Further statistical justification of this choice is offered in section 5.3.1. The decisions on what and how many other attributes should be included in our synthetic data and the ordering of their generation are guided by the following main considerations:

(a) the perceived importance of a variable in 'determining' others: some characteristics are very important in determining others and thus need to be assigned at an early stage; whereas others are more dependent and can be introduced latter (Birkin and Clarke, 1988).

(b) the availability of suitable local data linking a variable with those already created: a variable may be thought of as important, but without local constraints to link it with existing variables estimates are unlikely to be accurate.

(c) data quality: the published small area tabulations have been modified and some only contain a 10% sample of all census returns.

(d) cost: every additional variable included will increase the time for programming and testing.

4.3.1 Variables and their ordering

If we accept that given the location of a household the characteristics of household head are important factors in generating other variables, then we can start with analysing what variables are good predictors of a person being a head (or non-head). Many variables within the census may be relevant to the analysis, but only those contained in the census tabulations that are directly linked with household heads are under consideration. With the aid of Metac91, a meta-database about 1991 Census table contents (see Williamson, 1993; Williamson *et al.*, 1995), we found five SAS tables plus two LBS tables are of greatest relevance. They are S39, S49, S51, S86, S90, L45 and L50. Figure 5 shows these tables (in dark fill) and the variables appearing in the tables. There are nine variables all together:

Age	Economic primary position
Sex	Socio-economic group
Marital status	Social class
Tenure	Country of birth
Ethnic group	

In Figure 5 we plot three sets of age, sex, and marital status, which represent three population groups: household heads, household residents, and all residents. There are two reasons for this. First, as shown in Figure 5, few variables are directly linked with

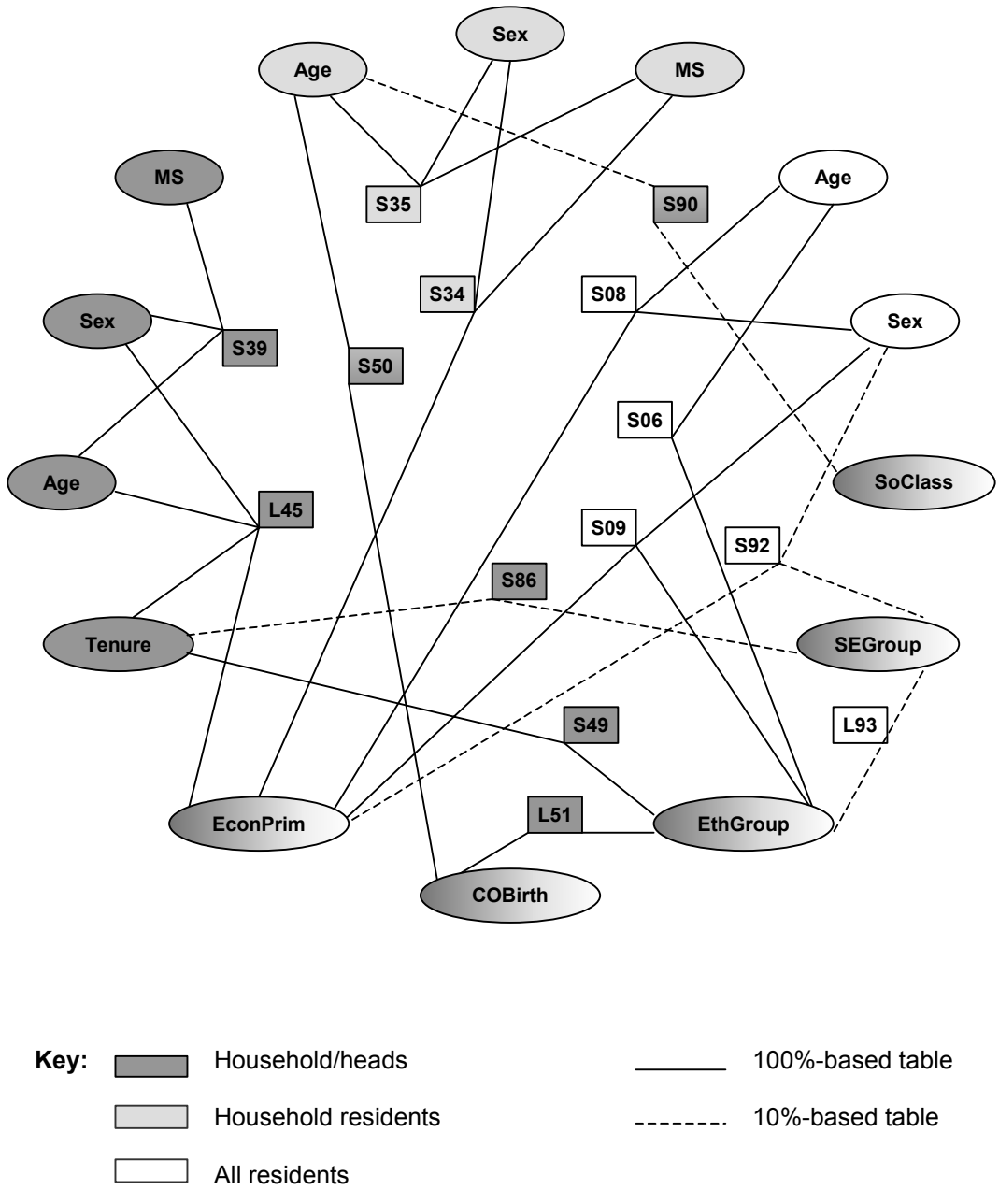


Figure 5 Census tables and the links between selected variables

the demographic characteristics of household head, but some tables connect one of the household head characteristics with the demographic characteristics of the parent population group such as household residents. For example, S50 gives the breakdown of age of household residents by country of birth of household heads. Second, by separating the three population groups we obtain a clearer picture of the relevant tables and linkages between variables, which has proven to be a useful visual aid in identifying suitable constraining tables.

These nine variables are all potential predictors of a person being a head (or non-head). Because the dependent variable is dichotomous (head or non-head) we can use logistic regression analysis to identify those that provide the best predictors of headship. The data allowing us to do so are drawn from the SAR. The whole SAR contains a very large sample. For the purpose of the current analysis a 10% random sample of of the household records was extracted. Ten variables, nine described above plus relationship to household head (*relat*), were retained for all individuals aged 16 and over. The SPSS forward selection logistic regression algorithm was then used to single out the key predictors of headship.

One of the problems of using logistic regression is that there is no commonly accepted measure of 'goodness of fit'. We used two SPSS outputs: (a) the classification table, which compares model predictions to the observed outcomes; and (b) the –2 log likehood (-2LL) values. Table 2 summarises the results of our model. It shows the variables entered into the model at each step. For each step the correct prediction percentage of the model from the classification table and the improvement in –2LL achieved are reported.

The first variable selected by is sex, suggesting that this is the single most important predictor of headship out of the nine available in published census ED outputs. Just using sex, 79.1% of household heads and 71.8% of non-heads are correctly classified, with an overall correct prediction rate is 75.6%. The next variables selected by forward regression were marital status, followed by age. Having included these three variables the overall correct prediction rate reaches 84.7%. At the end of its run, the logistic regression analysis selected eight out of the nine available variables. The variable not selected was country of birth. As shown in Table 2, the correct prediction rate barely

Table 2 Summary results of the logistic regression model

Step	Variable entered *	(a) Correct predictions (%)			(b) -2 log likelihood
		Head	Non-head	Overall	
1	Sex	79.1	71.8	75.6	34514
2	MStatus	88.2	69.9	79.3	29889
3	Age	84.7	84.7	84.7	29103
4	Tenure	84.6	85.1	84.9	28587
5	SEGroup	84.4	84.2	84.3	28065
6	EconPrim	84.5	84.0	84.3	27775
7	EthGroup	84.5	84.0	84.3	27741
8	SoClass	84.5	84.0	84.3	27718

* COBirth is not included in model because statistically significant.

changes after step three, although the –2LL continues to decrease as more variables are included in the model.

The results provide clear evidence that sex, marital status and age are the three main predictors of headship. Country of Birth turns out to be statistically non-signficant as a predictor, having taken into account the other variables available, and may be discarded. Social Class, although statistically significant, is also discarded as a predictor of headship at this stage. as a potential modelled head of household attribute. It was the last (and by implication least important) predictive variable identified via forward regression, is only 10% coded, has significant conceptual overlaps with socio-economic group, and has fewer links with other target household head attributes.

Having generated household head's age, sex and marital status (neatly cross-classified in S35), an obvious step is to combine the data in S35 and S39 to generate non-heads by age, sex and marital status, and to disaggregate heads' ages into finer categories. After this the remaining household head attributes of tenure, SEG, economic position and ethnic group are synthetically reconstructed in an order determined primarily by data availability.

From Figure 5 we can see that of the four head's attributes to be added, only two are linked with the household head attributes of age, sex and marital status. S34 gives economic position by sex and marital status for household residents and S08 gives economic position by age and sex for all residents. At ward level, L45 cross-tabulates the age, sex and economic position of household heads and tenure. Economic position is selected as the first variable to be generated after age, sex and marital status, because we have more data liking this variable with age, sex and marital status. Economic position is added to heads and non-heads alike. After economic position we generate tenure, followed by ethnic group of household head and finally socio-economic group of household head. Socio-economic group is generated after ethnic group because S86 is inflated from 10%-based table and not as reliable as S49. Figure 6 shows the sequence of steps in population generation and Table 3 reports the modelled attributes and their details.

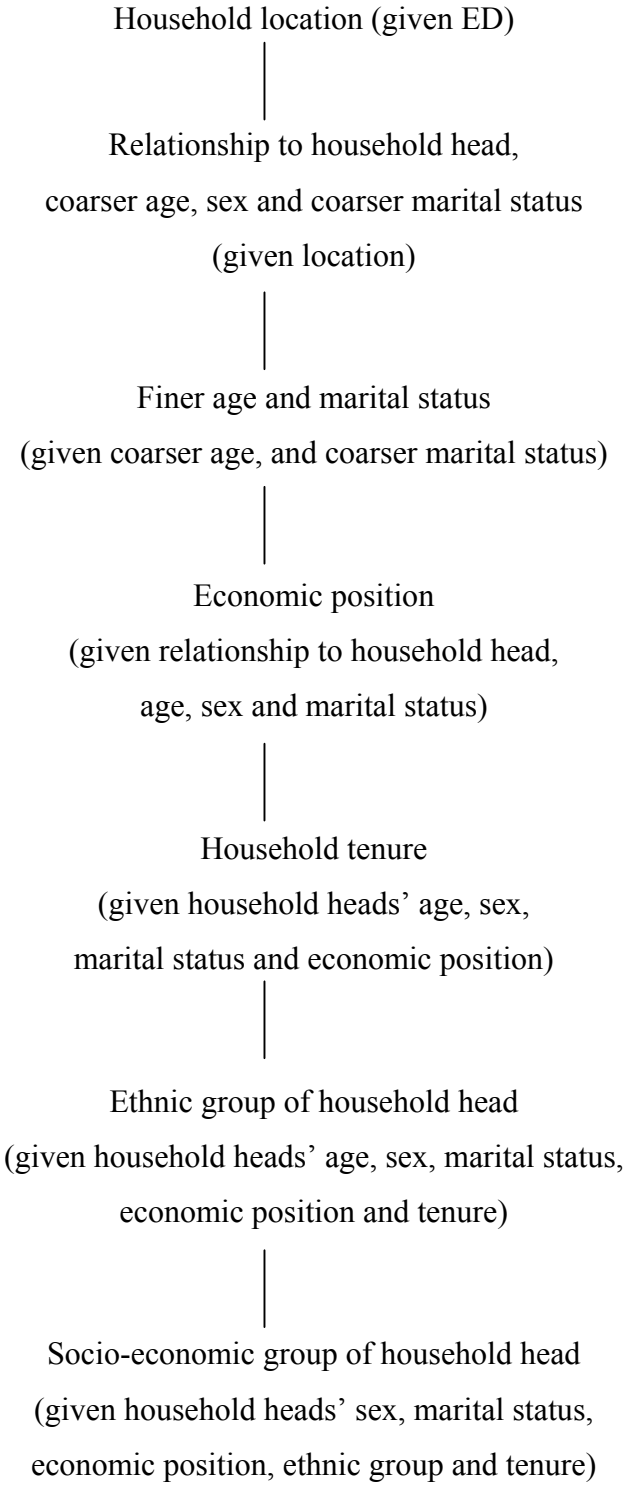


Figure 6 Sequence of steps in population reconstruction

Table 3 Attributes and their details

Location

Enumeration district

Relat*: relationship to household head* (2)

1 Head of household
2 Non-head of household

Age*:* (75)

16, 17, …, 89, 90+

Sex*:* (2)

1 Male
2 Female

MStatus*: marital status* (4)

1 Single
2 Married
3 Widowed
4 Divorced

Tenure*:* (7)

1 Owner occupied-owned outright
2 Owner occupied-buying
3 Rented privately-furnished
4 Rented privately-unfurnished
5 Rented with a job or business
6 Rented from a housing association
7 Rented from a local authority or new Town

EcomPrim*: primary economic position* (10)

1 Employees-full time
2 Employees-part time
3 Self employed-with employees
4 Self employed-without employees
5 On a government scheme
6 Unemployed
7 Students
8 Permanently sick
9 Retired
10 Other economically inactive

EthGroup*: Ethnic group of household head* (10)

1 White
2 Black Caribbean
3 Black African
4 Black Other
5 Indian
6 Pakistani
7 Bangladeshi
8 Chinese
9 Other groups - Asian
10 Others

SEGroup*: Socio-economic group of household head* (18)

1 Employers and managers in large establishments
2 Employers and managers in small establishment
3 Professional workers - self-employed
4 Professional workers - employees
5 Ancillary workers and artists
6 Foremen and supervisors - non-manual
7 Junior non-manual workers
8 Personal service workers
9 Foremen and supervisors - manual
10 Skilled manual workers
11 Semi-skilled manual workers
12 Unskilled manual workers
13 Own account workers (other than professional)
14 Farmers - employers and managers
15 Farmers - own account
16 Agricultural workers
17 Members of armed forces
18 Inadequately described and not stated occupations

Figure in parentheses indicates the number of categories

In theory, other variables, such as household composition and size, can be added sequentially given the household head characteristics that have been created. But in practice ward and ED level tables provide very poor linkages between household composition and head of household characteristics. Instead, to proceed further, crucial links would have to be made drawing upon national or regional distributions. However, the variables listed in Table 3 cover the key population characteristics and are sufficient for the purpose of this paper, which is to compare the two main approaches to the generation of small-area population microdata.

4.3.2 Population reconstruction

We now turn to detail of the population reconstruction process. Following the framework shown in Figure 6, the synthetic reconstruction model is divided into six steps (steps 0 to 5). The main task for each step is to create a joint distribution or conditional probability for a given ED so that a new variable can be added. As discussed in Section 4.2.1 the probabilities required are derived using a three-level estimation procedure. For each step of the reconstruction process Table 4 reports: the joint distribution at national level derived from the SAR, the ward-level constraining tables used, the variables involved including their number of categories, and the target ward-level joint distribution. In a similar way, Table 5 reports for each step the ED level constraining tables used and the target ED-level joint distribution.

Step 0: Generate an initial population

To begin with, we create a population of adults with following characteristics: relationship to head of household (head and non-head), coarse age, sex, coarse marital and location (ED). The number of household heads in the ED is identified by S71. S39 and S35 give the breakdown of the number by age, sex and marital status for household heads and household residents respectively. S35 has finer grouping of age than S39 (see Table 5, step 1), so it is aggregated to the S39 format. The difference of the two is the estimated distribution of non-heads by age, sex and marital status. Note that all of the constraining tables are subject to an adjustment procedure, as described in Section 4.2.1. From this we can create a sample with the four attributes: relationship to head of household, coarse age (7 classes: 16-29, 30-44, 45-59, 60-64, 65-74, 75-84, 85+), sex and

Table 4 Calculating ward level joint distributions

	Tables used	Variables and population group	Relat 2	Age 75	Sex 2	MStatus 4	EconPrim 10	Tenure 7	EthGroup 10	SEGroup 18
STEP 1										
	SAR1	Relat-Age-Sex-MStatus	**2**	**75**	**2**	**4**		National level joint distribution		
	L39	HOH: Age-Sex-MStatus		**9**	**2**	**3**				
	L35	RinH: Age-Sex-MStatus		**17**	**2**	**4**		Ward level constraints		
	L38	RinH: Age-Sex		**75**	**2**					
	WQ1	Relat-Age-Sex-MStatus	**2**	**75**	**2**	**4**		Ward level joint distribution		
STEP 2										
	SAR2	Relat-Age-Sex-MStatus-EconPrim	**2**	**17**	**2**	**2**	**10**			
	L45	HOH: Age-Sex-EconPrim		**4**	**2**		**3**			
	L34	RinH: Sex-MStatus-EconPrim			**2**	**2**	**10**			
	L08	AR : Age-Sex-EconPrim		**17**	**2**		**10**			
	WQ12	Relat-Age-Sex-MStatus	**2**	**17**	**2**	**2**				
	WQ2	Relat-Age-Sex-MStatus-EconPrim	**2**	**17**	**2**	**2**	**10**			
STEP 3										
	SAR3	HOH: Age-Sex-MStatus-EconPrim-Tenure		**4**	**2**	**2**	**10**	**7**		
	L45	HOH: Age-Sex-EconPrim-Tenure		**4**	**2**		**3**	**4**		
	L42	HOH: Tenure						**7**		
	WQ23	HOH: Age-Sex-MStatus-EconPrim		**4**	**2**	**2**	**10**			
	WQ3	HOH: Age-Sex-MStatus-EconPrim-Tenure		**4**	**2**	**2**	**10**	**7**		
STEP 4										
	SAR4	HOH: Age-Sex-MStatus-EconPrim-Tenure-EthGroup		**4**	**2**	**2**	**10**	**5**	**10**	
	L49	HOH: Tenure-EthGroup						**5**	**10**	
	L06	AR: Age-Sex-EthGroup		**4**	**2**				**10**	
	L09	AR: Sex-EconPrim-EthGroup			**2**		**10**		**10**	
	WQ34	HOH: Age-Sex-MStatus-EconPrim-Tenure		**4**	**2**	**2**	**10**	**5**		
	WQ4	HOH: Age-Sex-MStatus-EconPrim-Tenure-EthGroup		**4**	**2**	**2**	**10**	**5**	**10**	
STEP 5										
	SAR5	HOH: Sex-MStatus-EconPrim-Tenure-EthGroup-SEGroup			**2**	**2**	**2**	**5**	**4**	**18**
	L86	HOH: Tenure-SEGroup						**5**		**18**
	L92	AR: Sex-EconPrim-SEGroup			**2**		**2**			**18**
	L93	AR: EthGroup-SEGroup							**4**	**9**
	WQ45	HOH: Sex-MStatus-EconPrim Tenure-EthGroup			**2**	**2**	**2**	**5**	**4**	
	WQ5	HOH: Sex-MStatus-EconPrim Tenure-EthGroup-SEGroup			**2**	**2**	**2**	**5**	**4**	**18**

Notes: HOH - Head of Household; RinH - Residents in household; AR - All residents; see Table 3 for other variable codes
Figures indicate the number of categories

Table 5 Calculating ED level joint distributions

	Tables used	Variables and population group	Relat 2	Age 75	Sex 2	MarStatt 4	EconPrim 10	Tenure 7	EthGroup 10	SEGroup 18
STEP 1										
	WQ1	Relat-Age-Sex-MStatus	**2**	**75**	**2**	**4**		Ward level joint distribution		
	S39	HOH: Age-Sex-MStatus		**7**	**2**			ED level constraints		
	S35	RinH: Age-Sex-MStatus		**17**	**2**	**2**				
	EQ1	Relat-Age-Sex-MStatus	**2**	**75**	**2**	**4**		ED level joint distribution		
STEP 2										
	WQ2	Relat-Age-Sex-MStatus-EconPrim	**2**	**17**	**2**	**2**	**10**			
	S86	HOH: EconPrim					**2**			
	S34	RinH-Sex-MStatus-EconPrim			**2**	**2**	**10**			
	S08	AR : Age-Sex-EconPrim		**9**	**2**		**10**			
	EQ12	Relat-Age-Sex-MStatus	**2**	**17**	**2**	**2**				
	EQ2	Relat-Age-Sex-MStatus-EconPrim	**2**	**17**	**2**	**2**	**10**			
STEP 3										
	WQ3	HOH: Age-Sex-MStatus-EconPrim-Tenure		**4**	**2**	**2**	**10**	**7**		
	S42	HOH: Tenure						**7**		
	EQ23	HOH: Age-Sex-MStatus-EconPrim		**4**	**2**	**2**	**10**			
	EQ3	HOH: Age-Sex-MStatus-EconPrim-Tenure		**4**	**2**	**2**	**10**	**7**		
STEP 4										
	WQ4	HOH: Age-Sex-MStatus-EconPrim-Tenure-EthGroup		**4**	**2**	**2**	**10**	**5**	**10**	
	S49	HOH: Tenure-EthGroup						**4**	**4**	
	S06	AR: Age-EthGroup		**3**					**10**	
	S09	AR: Sex-EconPrim-EthGroup			**2**		**3**		**4**	
	EQ34	HOH: Age-Sex-MStatus-EconPrim-Tenure		**4**	**2**	**2**	**10**	**5**		
	EQ4	HOH: Age-Sex-MStatus-EconPrim-Tenure-EthGroup		**4**	**2**	**2**	**10**	**5**	**10**	
STEP 5										
	WQ5	HOH: Sex-MStatus-EconPrim-Tenure-EthGroup-SEGroup			**2**	**2**	**2**	**5**	**4**	**18**
	S86a	HOH: Tenure-SEGroup						**5**		**18**
	EQ45	HOH: Sex-MStatus-EconPrim Tenure-EthGroup			**2**	**2**	**2**	**5**	**4**	
	EQ5	HOH: Sex-MStatus-EconPrim Tenure-EthGroup-SEGroup			**2**	**2**	**2**	**5**	**4**	**18**

Notes: HOH - Head of Household; RinH - Residents in household; AR - All residents; see Table 3 for other variable codes
Figures indicate the number of categories

coarse marital status (2 classes: single/widowed/divorced and married), which match known distributions exactly.

Step 1: Disaggregate the sample into finer age and marital status

The next step is to disaggregate the sample generated in step 0 into target groupings (i.e., breaking age down into single year groups and marital status into single, married, widowed, and divorced groups). A joint distribution of four variables with the target groupings is derived from the SAR, which is denoted by SAR_1. Ward-level constraints are L39, L35 and L38 (Table 4, step1). Using the IPF procedure we obtain a ward-level four-variable joint distribution WQ_1, which is, in turn, scaled down to fit the ED level constraints of S39 and S35 using IPF. The result is an ED-level joint distribution EQ_1 (Table 5, step 1). So we can calculate the conditional probability distributions of the finer age and finer marital status given sample's relationship to head of household, coarse age, sex and coarse marital status. Using modified Monte Carlo sampling we disaggregate our sample into the target variable details.

Step 2: Generate economic position

The process of adding variables is more complex than data disaggregation. In step 2 we wish to estimate the economic position for our sample, which is assumed to be conditionally dependent upon the 'known' variables generated in step 1. At step 1 the estimated joint distribution is an array with $2\times75\times2\times4 = 1200$ cells. Economic position has 10 categories. If we want create a full five-variable distribution using the finest possible variable categorisation we would end up with an array containing 12,000 cells, the majority of which would be empty. An efficient solution is to divide the sample into subgroups. For example, we know that the majority of females aged over 60 are retired, so we do not need to break age down into single year groups to estimate economic position. However aggregating these people into one group may be too coarse, as we may wish to examine, for instance, the relationship between age and the economic position 'permanently sick'. When several variables are involved, the partition of the sample into subgroup is not straightforward. Moreover, as the process goes on and more we have to select a subset of all possible predictors to reduce complexity. Again, the

decision is guided by (a) analysing the relationships between known variables and the target one, and (b) the availability of the data.

The target (dependent) variable has more than two categories ruling out the use of logistic analysis for identifying the most suitable predictors. The Quest decision-tree algorithm (Loh and Shih, 1997), as implemented in SPSS AnswerTree, provides a useful alternative tool, which allows for the analysis of multi-category and continuous variables. It is a tree-based analysis, which uses chi-squared or F statistics to select predictors, with the results presented in a decision tree map. At the top of the tree is the target variable, with the next level down showing the first predictor selected by the model, which is split into several nodes. As the tree grows into many levels we have a map that shows the combined effect of the selected variables upon the target variable. Figure 7 shows a simple two-level tree map of for the prediction of economic position using the head of household variables already modelled in step 1. The data used for this analysis are the same sample as we used for the logistic model. The first predictor selected is sex, which is split into male and female. The root node is a tabulation of the target variable. Comparing the two tabulations at the first level we can see different employment patterns between men and women. Proportionally more men work full time (51%) than women (26%), but more women work part time (16.5%) than men (1.7%). At level two each node is split into head and non-head of household. Figure 8 shows the result when we let the tree grow a further two levels. For simplicity we ignore the tabulations at each node and concentrate simply on the variables and their division at each level. Age is the predictor selected by the model at level three, but the partition is different for different nodes. At level four marital status and age (further disaggregation) are included in the model.

To see how well the model does at predicting economic position we can examine the risk summary, which compares the tree's assignment of economic position with the position actually recorded. Table 6 shows the risk estimate at each level. The number of levels is five. The risk estimate gives the proportion of cases classified incorrectly. We can see the misclassification rate of the model decreases as the number of levels increases, but the improvement is trivial at level five. Although decision trees can grow into many levels, further splits may add little to our understanding of the problem because the subsequent splits deal with small numbers of cases.

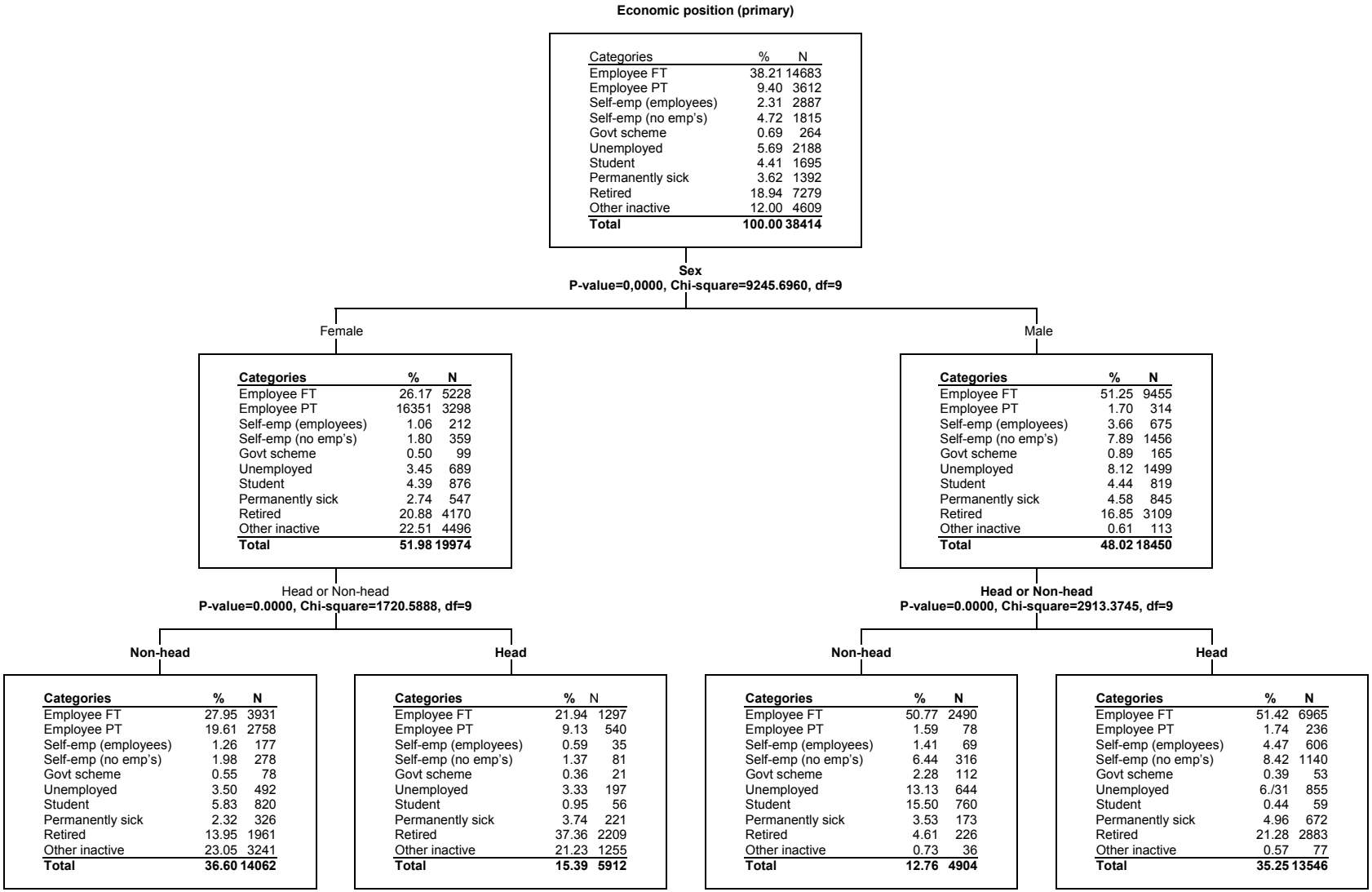


Figure 7 Two-level decision-tree map

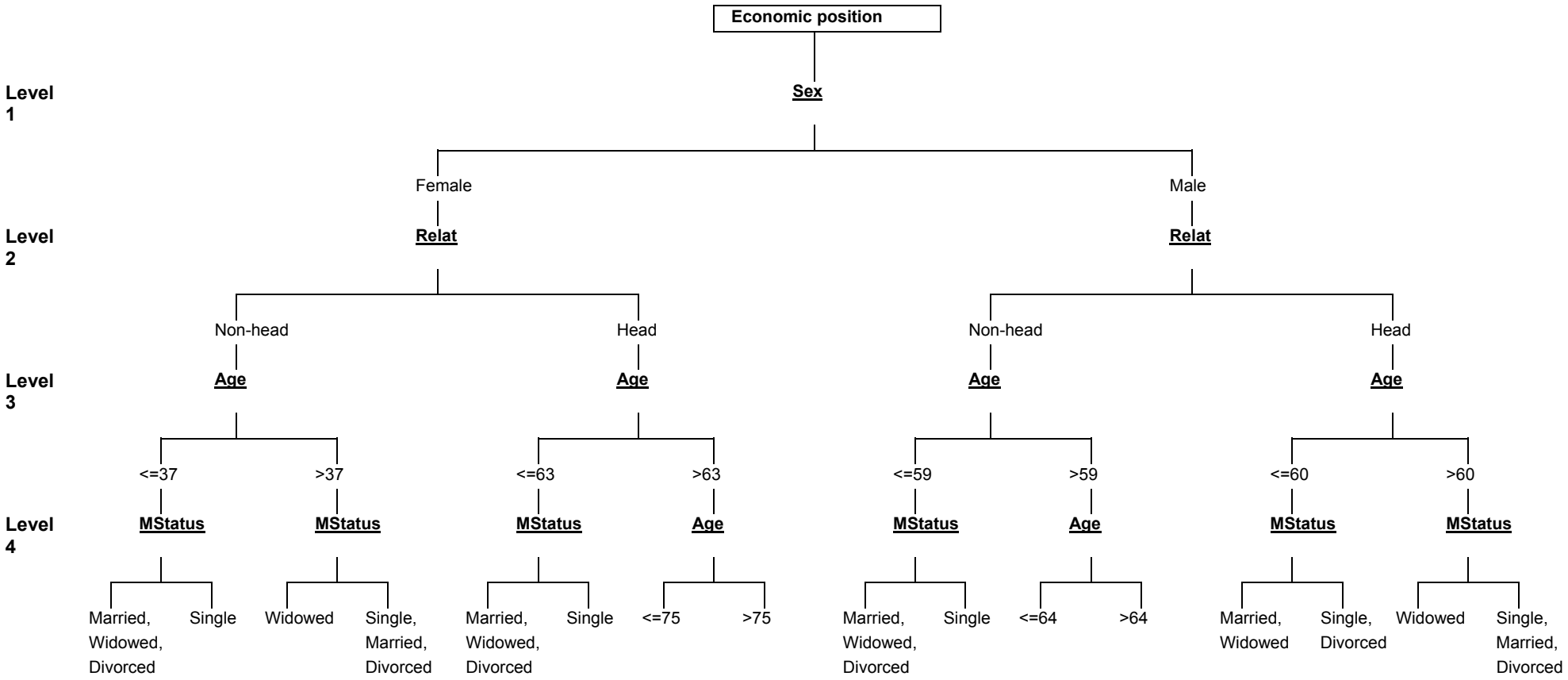


Figure 8 A full decision tree map identifying the determinates of economic position

Table 6 Risk estimated by the tree-based model

Level of split	Variables	Risk estimated
1	Sex	0.618
2	Relationship to head	0.594
3	Age	0.490
4	Marital status or Age	0.484
5	Marital status or Age	0.479

The dependent variable is economic position.

The results suggest we need all four known variables to estimate the economic position of our synthetic population. Even including all these variables the misclassification rate of the model is still high (about 48%), which is why the local constraints play an important role. When deciding how to reduce the size of the joint distribution, the golden rule is that we do not reduce the known local information. From Table 4 (step 2) we can see that L08 links age (17 classes) with economic position and L34 links marital status (2 classes) with economic position. So at step 2 age is aggregated into 17 classes, whilst marital status is aggregated into 2 classes. Figure 8 suggests it is better to separate single from divorced/widowed. This can be compromised by the finer age scheme.

A joint distribution of five variables, SAR_2 ($2\times17\times2\times2\times10$), is derived from the SAR. Ward-level constraints are L45, L34 and L08 (Table 4, step 2). L45 gives the breakdown of age by sex by economic position by tenure for household heads. Only the first three variables are used at this step. ED-level constraints are part of S86, which gives the proportion of economically active and economically inactive household heads, S34 and S08.

It should be noted that tables L08 and S08 are for all residents and are scaled down in this step to represent residents in household only. This could cause a net error in table counts. These tables are also very large (S08 contains 180 cells), so the exact table counts are not reliable. On the other hand, ignoring these tables would certainly lose some local information. An innovation is adopted at this point, which is to treat these tables as quasi-constraints. When using the IPF procedure quasi-constraints are only used for a few iterations (e.g., 5, 10). The final estimates should retain some of the interaction patterns of the quasi-constraints but will not necessarily match their counts.

The concept of quasi-constraint could also help reduce the complexity of IPF when dealing many variables. For example, at this step we need to estimate a full five-variable joint distribution at ward-level. The constraining tables mentioned above (L45, L34 and L08) are only those containing economic position. We should also include tables L39 and L35, which have been used in step 1, so that the estimated joint distribution fits all known constraints. This certainly increases the complexity of the procedure. When dealing with more variables we may find it too cumbersome to operate. We suggest using the distribution created at the previous step instead of the constraints already used. As

shown in Table 4 (step 2), we use WQ_{12} as a constraint in step 2, which is obtained by aggregating WQ_1 to meet the scheme of the first four variables in SAR_2. Consequently, WQ_{12} fits both L39 and L35. However, WQ_{12} is a four-dimensional array and the cell counts are the maximum likelihood estimates. We can treat this as a quasi-constraint if its existence causes the problem of convergence in IPF. Otherwise, it can be used as a normal constraining table. ED level constraints can be treated in a similar way. With the use of IPF we obtain a ward level and ED level distributions of five variables, WQ_2 and EQ_2, resulting in the conditional probability distributions of economic position given sample's relationship to head of household, coarse age, sex and coarse marital status. At the end of step 2 each individuals in our sample has been assigned a category of economic position.

Step 3 to 5: Generate tenure, ethnic group and socio-economic group for household heads

In Steps 3 to 5 we generate household tenure, ethnic group and socio-economic group for household heads. These are head of household characteristics, so the existing non-head attributes (age, sex, marital status and economic position) are not involved in the processes. At each step, the Quest decision-tree algorithm has been used to help select the most appropriate predictors and decide their levels of aggregation. For estimating tenure and ethnic group all known characteristics for household heads are used (see Tables 4 and 5, steps 3 and 4). For estimating socio-economic group one variable (age) is eliminated. As a result, we need to create an array of six dimensions at steps 4 and 5. Using the technique described above this can be achieve without too much difficulty, because the constraints for each step are only the tables linking the target variable with the predictors and the joint distribution of the predictors derived from the previous step.

In total thirteen LBS tables and nine SAS tables have been used in the population reconstruction process. These tables plus the joint distributions for each step derived from the SAR act as the inputs of Pop91SR. The output of the model is a synthetic dataset with all the variables described in Table 3 for a given area. An assessment of the quality of this dataset will be presented in Section 6.

5. The combinatorial optimisation model (Pop91CO)

5.1 Model components and method employed

Pop91CO is the latest version of a program suite used previously for the creation of synthenetic microdata using combinatorial optimisation approach (see Williamson, 1996; Williamson *et al.*, 1998). It consists of three sets of programs. The first set of programs is designed to extract SAS table data and convert them into carefully ordered table vectors, which act as the constraints for household combinations. The second program suite assigns each household and individual in the SAR with a set of values; each value indicates the cell number in a given table vector to which each individual/household relates. This speeds the process of casting SAR data into SAS look-alike tables to enable statistical comparison of population distributions between data recorded in SAS and SAR formats. The third and main program suite is concerned with the evaluation and selection of combinations of households from the SAR that best fit constraining tables. It includes a number of subroutines which select an initial set of households, iteratively evaluate the effects of replacing one of the selected households until a satisfactory fit is reached, and report the results, respectively.

The combinatorial optimisation approach is relatively simple compared with the synthetic reconstruction approach. It attempts to fit all the selected constraints simultaneously. However, to achieve a satisfactory fit across all the constraints is not an easy task. The quality of the resulting synthetic dataset is likely to be affected by the these factors: the size of the sample used as a parent population; the constraints used to guide the household selection; the method of combinatorial optimisation; the selection criterion; the computer resource used; and the divergence of each small area's characteristics from norm.

(1) *The size of the sample used as a parent population*. In the combinatorial optimisation approach it is assumed that the population characteristics in a small area can be reassembled by a set of households drawn from a known sample. The larger the sample size, the more possible combinations of households exist and the better the fit is likely to be. The SAR contains around 215,000 household records. It is might well be possible to find household combinations from such a large pool that match local population characteristics of various types. But the solution space is

extremely large. One might use a smaller sample, such as region-specific SAR instead of whole SAR. It is, however, *a priori*, not clear whether using region-specific SAR would generate the same level of fit as using whole SAR. Further consideration is given to this question in the next section. Even using the whole SAR there is no guarantee that every type of household will be represented, since it is a 1% sample. When a perfect match cannot be found (or could not be discovered in the time allowed), the result is likely to be a population more akin to the national (SAR) average than actually exists in the area being considered.

(2) *The constraints used to guide the household selection*. The number of constraining tables adopted and the variables involved in these tables will affect the resulting output. Previous work (Voas and Williamson, 2000a) suggests that synthetic microdata generally produce a poor fit to tabulations of variables not used as constraints. Using a different set of constraints would produce different results. In general, the more constraints used the better the synthetic dataset. But every additional table included will increase computing time, as more iterations will be required to achieve a given level of fit. At the experimental stage, a set of eight SAS tables was used to judge the fitness of a household combination (Williamson *et al*., 1998; Voas and Williamson, 2000a). These are the first eight tables listed in Table 7. The selection of these tables was guided by the desire to include an equal number of household and individual-level tabulations, covering as wide a range of census variables as possible in as few tables as possible. It was also partly motivated by assessments of the relative interest of particular topics to researchers.

It is quite possible that a better set of constraints could be chosen. The primary objective of this paper, however, is to compare the effectiveness of the two main approaches to generation of synthetic microdata. To ensure comparability, the approaches should adopt the same set of constraints. Of the nine SAS tables used in the synthetic reconstruction process only two are included in the original list of constraints for the combinatorial optimisation model Hence seven more tables have been added to the list of constraining tables, resulting in a total of fourteen possible constraints (see Table 7). A switch is assigned to each table, and users can select a set of the constraints of most interest. For the purpose of this paper, the nine SAS tables used in building Pop91SR have been selected to guide the selection of

household combinations. As the 10%-based table counts for EDs are known to be unreliable (Voas and Williamson, 2000a), we have already substituted them with the revised the table counts for S86 estimated during the synthetic reconstruction process.

(3) *The method of combinatorial optimisation*. In the early stages of model development, considerable attention was devoted to identifying the best methods of combinatorial optimisation. Williamson *et al.* (1998) tested three techniques of combinatorial optimisation: hill climbing, simulated annealing and genetic algorithms. The results suggested that modified simulated annealing (a hybrid of hill climbing and simulated annealing) stood out as the best solution (in term of the greatest average reduction in total absolute error). Details concerning the assessment of these techniques may be found in the above study. To summarise, in both hill climbing and simulated annealing algorithms an initial combination of households are selected from the SAR. Subsequently a household from the combination and a possible replacement from the SAR are randomly selected. In hill climbing the replacement will be made only if the swap improves the fit, whilst in simulated annealing some swaps are accepted even if they lead to a moderate degradation in performance, in order to allow the algorithm to backtrack from suboptimal solutions. The probability of this 'retrograde swap' decreases with the increase of the number of successful replacements and is determined by two parameters: the starting 'temperature' and the ratio of initial temperature to number of replacements. When the starting temperature is set low the probability of a retrograde swap occurring soon declines towards zero, at which point the behaviour of simulated annealing becomes much more like that of straightforward hill climbing. It is this latter approach that we describe as 'modified simulated annealing'.

(4) *The selection criterion*. The evaluative statistic used in the iterative fitting process directly affects household selection. Previous model variants used overall total absolute error as the selection criterion. The choice was guided by the desire to reduce computing time since it is very simple to calculate. The disadvantage is that TAE is a relatively crude measure, taking no account of error relative to the size of a cell count.. Using this measure, as reported in Williamson *et al*. (1998), it can frequently occur that even though the overall fit seems good, one or more tables do

Table 7 SAS tables included by Pop91CO

No.	SAS tables	Variables	Used for comparison of Pop91SR and Pop91CO
1	S01	Resident status / Sex	
2	S14	Long-term illness / Age / Economic activity	
3	S22	Household size / Number of rooms / Tenure	
4	S29	Dependants	
5	S35	Age / Sex / Marital status	✓
6	S42	Household composition / Tenure	✓
7	S74	Occupation / Age / Sex	
8	S86	Socio-economic group of household head / Tenure	✓
9	S06	Age / Ethnic group	✓
10	S08	Age / Sex / Economic position	✓
11	S09	Sex / Economic position / Ethnic group	✓
12	S34	Sex / Marital status / Economic position	✓
13	S39	Age / Sex / Marital status of household head	✓
14	S49	Ethnic group of household head / Tenure	✓

not fit particularly well. Voas and Williamson (2000a) suggested that using global measures will always produce closer matches for tables reflecting distributions similar to the overall population's, at the expense of those with more divergent characteristics. This problem, however, may be particularly associated with the use of TAE. As we will see in section 5.2, adopting a new selection criterion both 'normal' and 'abnormal' tables will benefit from potential replacements, leading to significant improvements in resulting outputs.

(5) *The computer resource*. The process of combinatorial optimisation is iterative. In order to achieve a satisfactory fit hundreds of thousand or millions of evaluations are required for each ED. In fact, all the techniques and model designs are based around this theme: finding the best possible solution within the available time. Williamson *et al.* (1998) reported that for eight constraining tables 70 CPU seconds per ED are required to perform 500,000 evaluations on a powerful workstation. Due to advances in computer technology it is now possible to perform millions of evaluations on a desktop PC within a minute. As a result, we are now able to devote more time on 'hard-to-fit' area and include more tables as constraints in combinatorial optimisation.

(6) *The divergence of a small area's characteristics from norm*. The fit between constraining tables and synthetic microdata produced by combinatorial optimisation varies with location. The poorly fitting are typically those where observed statistics reflect a distribution very different to the national norm, and hence from that captured in the SAR. A challenge is how to fit these atypical areas. Voas and Williamson (2000a) developed a sequential fitting procedure in order to improve the accuracy of model outputs. Using this procedure they found every constraining table can be satisfied (i.e., SSZ not exceeding the critical value). However, further examination of this technique indicates that with the sequential fitting the SSZ statistics of abnormal tables can be reduced to an acceptable level, but at the expense of increasing overall TAE. In some cases the increase is substantial. For example, testing on ED DAGF12 in University ward of Leeds city, with the original simultaneous fitting procedure one table is always 'non-fitting' (NFT) (see next section). With sequential fitting all the tables can be fitted, but the overall total absolute error increases by more than 100% compared with the result of the original fitting procedure. It is also

difficult to apply the sequential fitting technique for generating large area microdata, because, as mentioned in Section 2.3, the ordering of tables to be fitted is area-specific.

Pop91CO has employed the modified simulated annealing approach but dropped the sequential fitting procedure. We have discussed the factors that may affect the modelling results. It is argued that the problem of poorly fitted tables partly results from the use of TAE in the iterative process, and that the sequential fitting procedure to cope with the problem is unsatisfactory. To solve these problems we propose an alternative selection criterion. This new measure is presented in the next section together with a number of techniques designed to improve the accuracy and consistency of model outputs.

5.2 New developments in Pop91CO

5.2.1 Selection criterion

The test statistic used in the iterative fitting process is vital in a combinatorial optimisation model. Previous versions have used overall total absolute error as the selecting criterion, defined as

$$Overall\ \ TAE\ \ =\ \ \sum_k \sum_i \left| O_i^k - E_i^k \right| \qquad (1)$$

where O_i^k and E_i^k are the observed and expected counts respectively for *i*th cell of *k*th table vector. The drawbacks of this measure have already been discussed. In Section 3.2 we presented a new summary statistic, overall RSSZ (relative sum of squared Z scores), as the measure of global fit across several tables, based upon the work of Voas and Williamson (2001a). It is calculated by

$$Overall\ \ RSSZ\ \ =\ \ \sum_k \frac{SSZ^k}{C^k} \qquad (2)$$

where SSZ^k is sum of squared Z scores for table *k*

$$SSZ^k\ \ =\ \ \sum_i Z_i^{k\,2} \qquad (3)$$

and C^k is the 5% χ^2 critical value for table *k*. Z_i^k is the Z score for *i*th cell of *k*th table vector, which is calculated as:

$$Z_i^k = \left(\frac{E_i^k}{N_e^k} - \frac{O_i^k}{N_o^k}\right) \Big/ \sqrt{\frac{1}{N_e^k}\frac{O_i^k}{N_o^k}\left(1 - \frac{O_i^k}{N_o^k}\right)} \quad (4)$$

where N_o^k and N_e^k are the observed and expected table totals respectively for table *k*.

During the iterative fitting process the expected totals may not be identical to the observed total, so the modified Z score (Z_m) is used. In a modified version the expected table total is replaced by the observed table total, hence

$$Z_{mi}^k = \left(E_i^k - O_i^k\right) \Big/ \sqrt{O_i^k\left(1 - O_i^k / N_o^k\right)} \quad (5)$$

When observed and expected table totals are the same, Z_m=Z.
From equations (2), (3) and (5) we obtain a modified overall RSSZ

$$Overall\ RSSZ_m = \sum_k \sum_i F_i^k \left(O_i^k - E_i^k\right)^2 \quad (6)$$

where F_i^k is a constant for each table cell,

$$F_i^k = \frac{1}{C^k O_i^k \left(1 - O_i^k / N_o^k\right)}, \quad O_i^k \neq 0 \quad (7)$$

or

$$F_i^k = \frac{1}{C^k}, \quad O_i^k = 0 \quad (8)$$

We can see from equation (6) that using the modified Z, everything except the difference in cell values is fixed for a given area. Comparing with equation (1), it is therefore quite feasible to use overall $RSSZ_m$ as an alternative test statistic in sampling iterations without greatly increasing computing overhead. At the later stage of sampling the expected and target totals will be highly similar as are $RSSZ_m$ and RSSZ. Therefore, it is consistent to use the overall $RSSZ_m$ as the selecting criterion and overall RSSZ as the measure to evaluate the final results.

Table 8 presents a comparison of the performance of the two alternative selection criteria, TAE and $RSSZ_m$. Three different types of EDs (DAFJ01, DAGF04 and DAGF12) are tested, which are selected from our test areas. Their distances from norm are shown in Figure 3. ED DAFJ01 in the Cookridge ward is a typical suburban area, and the position is relatively close to the national norm. EDs DADF04 and DADF12 are in the University

Table 8 Results from the use of TAE and RSSZm as the selecting criterion

Selection criterion:	TAE					RSSZm				
(A) ED DAFJ01 in Cookridge ward (198 households)										
Evaluations('000)	**TAE**	**RSSZ**	**NFT**	**NFC**	**CPU (s)**	**TAE**	**RSSZ**	**NFT**	**NFC**	**CPU (s)**
0	1438	124.60	9.0	117.8	0	1438	124.60	9.0	117.8	0
10	447	7.51	1.2	28	0	495	2.71	0	15.8	0
100	188	1.26	0	6.4	2	185	0.52	0	0.2	3
500	145	0.86	0	3.4	9	111	0.30	0	0	13
1,000	135	0.82	0	3.6	19	97	0.27	0	0	26
1,500	118	0.73	0	2.6	28	93	0.26	0	0	40
2,000	107	0.64	0	2.2	38	86	0.24	0	0	53
2,500	102	0.67	0	2.6	47	81	0.24	0	0	66
3,000	101	0.65	0	2.4	57	81	0.23	0	0	79
3,500	98	0.63	0	1.8	66	78	0.23	0	0	92
4,000	97	0.60	0	1.4	75	80	0.23	0	0	105
5,000	94	0.60	0	1.4	94	77	0.22	0	0	131
6,000	93	0.60	0	1.4	113	76	0.22	0	0	158
8,000	91	0.59	0	1.4	151	74	0.21	0	0	210
10,000	89	0.57	0	1.2	188	73	0.21	0	0	263
(B) ED DAGF04 in University ward (149 households)										
Evaluations('000)	**TAE**	**RSSZ**	**NFT**	**NFC**	**CPU (s)**	**TAE**	**RSSZ**	**NFT**	**NFC**	**CPU (s)**
0	1869	48.89	9.0	132.6	0	1869	48.89	9.0	132.6	0
10	880	12.11	5.6	67.0	0	853	8.14	3.0	59.2	0
100	364	4.11	0	21.8	2	359	1.88	0	4.6	3
500	320	3.57	0	18.8	9	236	0.98	0	0.6	13
1,000	275	3.07	0	14.8	19	206	0.81	0	0.4	26
1,500	248	2.72	0	13.2	28	200	0.75	0	0.4	39
2,000	240	2.58	0	13.8	38	190	0.70	0	0.2	53
2,500	233	2.35	0	13.0	47	185	0.67	0	0.2	66
3,000	227	2.22	0	12.0	57	178	0.65	0	0.2	79
3,500	222	2.16	0	11.8	66	173	0.62	0	0.2	92
4,000	219	2.18	0	10.4	75	177	0.62	0	0.2	105
5,000	216	2.16	0	10.4	94	171	0.60	0	0.2	131
6,000	213	2.11	0	10.6	113	162	0.58	0	0.2	158
8,000	207	2.03	0	10.0	151	160	0.56	0	0	210
10,000	201	1.98	0	9.8	188	153	0.53	0	0	263
(C) ED DAGF12 in University ward (191 households)										
Evaluations('000)	**TAE**	**RSSZ**	**NFT**	**NFC**	**CPU (s)**	**TAE**	**RSSZ**	**NFT**	**NFC**	**CPU (s)**
0	2642	106.81	9	164.6	0	2642	106.81	9	164.6	0
10	1542	35.15	8.6	116.8	0	1421	17.24	7.8	103.2	0
100	659	7.56	3.4	43.6	2	680	3.97	0	19.0	3
500	445	5.39	1.8	23.2	9	398	1.59	0	4.2	13
1,000	385	4.29	1.0	20.6	19	343	1.24	0	3.0	26
1,500	355	3.98	1.2	18.8	28	315	1.11	0	2.0	40
2,000	338	3.83	1.2	16.6	38	295	1.05	0	2.2	53
2,500	324	3.87	1.4	17.4	47	294	1.02	0	1.6	66
3,000	314	3.69	1.2	17.8	57	286	0.98	0	1.6	79
3,500	309	3.69	1.2	17.4	66	284	0.95	0	1.4	92
4,000	305	3.69	1.2	18.0	76	278	0.95	0	1.4	105
5,000	300	3.61	1.2	16.6	94	271	0.90	0	1.6	132
6,000	296	3.64	1.4	17.8	113	269	0.88	0	1.4	158
8,000	293	3.57	1.2	17.2	151	269	0.86	0	1.2	211
10,000	290	3.54	1.2	17.6	189	261	0.85	0	1.2	263

Figures are 5-run average
Total number of tables: 9; Total number of cells: 597
CPU time is central processing unit time in seconds on a 800MHz PC

ward; the former is far from the norm (outside the 98% percentile of the national distribution), and the latter is even more atypical (outside the 99.8% percentile of the national). For each ED combinatorial optimisation is run for a maximum of 10 million evaluations.

Figure 9 highlights the differences in TAE and RSSZ arising from use of the two alternative selection criteria. Figure 9a shows that for every ED the TAE statistics are significantly lower when using the new selection criterion. The greatest improvement brought about by using $RSSZ_m$, however, is in the reduction of RSSZ statistic (Figure 9b). Table 8 provides a more detailed set of results, based on 5-run averages, reporting measures of overall fit across all nine constraining tables plus associated run times (CPU seconds). A similar story is emerges whichever of the four test statistics presented, TAE, RSSZ, NFT and NFC, are examined, with the synthetic microdata created using the $RSSZ_m$ selection criterion out-performing that created using TAE.

The results in Figure 9 and Table 8 also reveal that the more an ED's distribution diverges from the national average, the greater the observed improvement is likely to be. In fact, with the use of $RSSZ_m$ abnormal tables no longer appear hard to fit. Take for example ED DAGF12 (an extremely atypical area). Using TAE as the measure of fit in sampling at least one constraining table does not fit, no matter how many evaluations have been performed. With the use of $RSSZ_m$ all the constraints are satisfied within 100,000 evaluations. At the cellular level, the gain of using $RSSZ_m$ is even greater, as only one or two cells out of 597 with Z score exceeding the critical values, while this figure would be more than 17 if TAE was used (Table 8c). The one possible negative of using $RSSZ_m$ as a selection criterion is that the run time is 26 CPU seconds per million evaluations per ED on an 800MHz PC, representing a 40% increase compared to TAE-based runs. But even if run time rather than number of evaluations is taken as the basis for comparison, the results in Table 8 show that $RSSZ_m$ remains a superior selection criterion, achieving better fit from considerably fewer evaluations.

5.2.2 Using region-specific SAR

Previous versions of the combinatorial optimisation model have selected households from the whole SAR (i.e. ignoring the geography of SAR records). The household SAR is

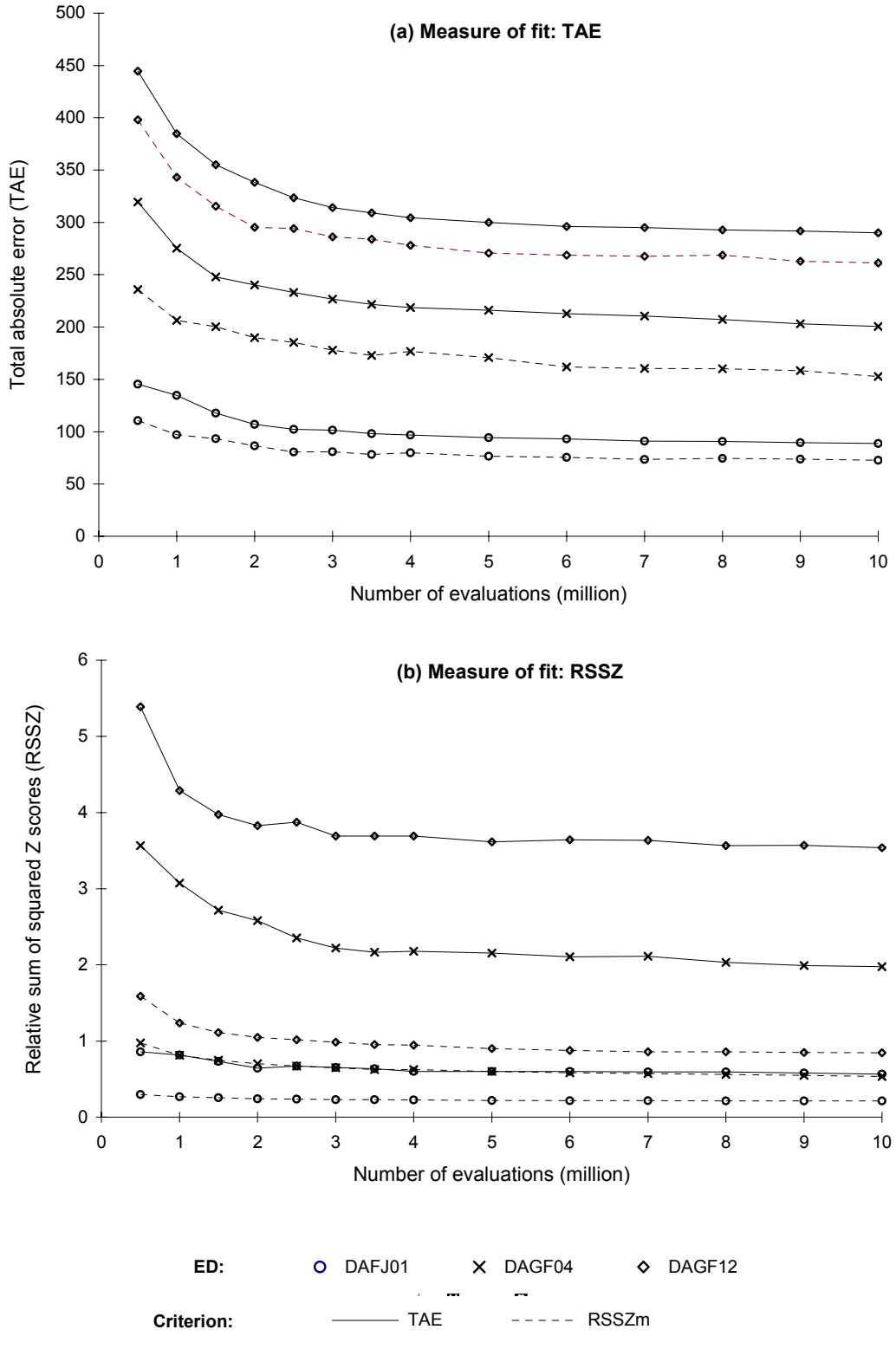
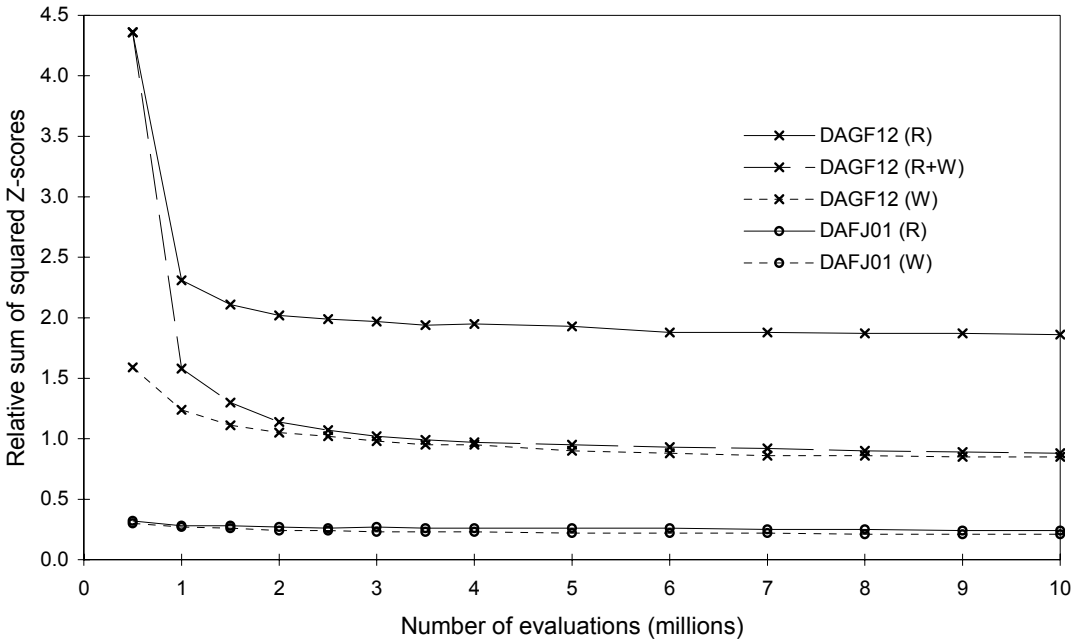


Figure 9 Performance of using alternative selecting criteria: TAE vs. RSSZm

spatially coded into twelve large areas (the Registrar General's Standard Regions, plus Wales and Scotland). The size of these regional samples vary greatly, ranging from about 8000 to 24000 households. *A priori*, it is not clear whether or not it is better to limit household selection to those households coming from the same region as the ED being synthesised. A major concern of using regional subsets of the SAR is that the solution space is reduced. However, the effect of using regional SAR might be very small for a typical ED, since there are still many households to select from. Problems may arise when several constraints are extremely atypical and no or few records from the regional SAR can match them.

Figure 10 highlights the difference between using a regional sub-set of the SAR (Scheme R) and whole SAR (Scheme W) tested on EDs DAFJ01 and DAGF12. Our test areas belong to the region of Yorkshire and Humberside. The SAR sample for this region contains approximately 19,000 household records. For ED DAFJ01 the difference between the RSSZ statistics of the two schemes is trivial, although the performance obtained using the whole SAR is slightly better. But for ED DAGF12 using only the relevant regional SAR performs significantly worse than using the whole SAR; the RSSZ statistic of Scheme R is almost twice of that of Scheme W. The results suggest that it is quite feasible to use regional subsets of the SAR for the majority of EDs where characteristics are close to the norm. But for atypical EDs limiting household selection to the regional SAR would significantly increase the error of estimation, and hence the whole SAR should be used. Even so, the use of regional SAR is to be preferred where possible, as the households in each subset may better reflect regional differences in household characteristics not constrained in the household selection process.

As a result, Pop91CO adopts a two-stage approach with respect to the usage of the SAR. In the first stage, the initial households are randomly drawn from the regional SAR that cover the EDs being synthesised and potential replacement elements are drawn from the same source. After a number of evaluations the results are analysed and the test statistic is compared with a pre-determined criterion. Currently, Pop91CO is set to assess whether or not the number of NFC is equal to zero after 200,000 evaluations. If the test is positive (i.e., not a single cell fails its Z-score test), it is assumed that the region-specific SAR is large enough to produce household combinations that match that ED's characteristics. All further potential replacement elements are therefore drawn from the regional SAR. If



R - region-specific SAR
W - whole SAR
R+W - region-specific and whole SAR

Figure 10 Comparing use of region-specific vs. whole SAR

the test is negative, in stage two the selection of potential replacement elements is switched from the regional to whole SAR. The results of the model after adopting this approach for the above two EDs are also plotted in Figure 10 (Scheme R+W). For ED DAFJ01 the result of Scheme R+W is the same as that of Scheme R, because the test to continue using the regional SAR was positive. For ED DAGF12, after switching the selection from regional to whole SAR, the RSSZ statistic soon declines towards the level of Scheme W, and after approximately three millions evaluations it is very close to the result of Scheme W. Using the whole SAR the final household combination in any given ED repolication contains less than 9% of households that come from the same region. Using the two-stage scheme the percentages of synthetic households come that from the relevant regional SAR are 100% and 14% for EDs DAFJ01 and DAGF12 respectively.

The use of regional SAR is guided by the intuition that the synthetic households drawn from a SAR region might fit better than those drawn from the whole SAR on topics that are not covered by constraining variables. Variables that are neither chosen to constrain the selection, nor are highly correlated with those that were chosen, will tend to be distributed according to the characteristics of the SAR. However, at such a coarse geography the distribution of region-specific SAR is unlikely to be much different from the national (the major exception perhaps being the South East and London). Therefore, the regional representation should not be over emphasised. Moreover, unconstrained variables generally do not fit well. Nevertheless, one might hope to lower the error at least to some extent by increasing the percentage of synthetic households from the region, subject to not degrading the model's performance on constraining variables.

5.2.3 Stopping rules

The results on the test of three EDs have shown that, using the new selection criterion, the estimated synthetic populations fit their constraining tables very well. As Table 8 shows, even for the extremely atypical area of ED DAFJ12, all the tables can be satisfied (no 'non-fitting' tables) within half a million evaluations (about 13 seconds computing time). Further iterations reduce the error at the cellular level, albeit at a gradually declining rate. Figure 9 shows that for ED DAFJ01 after half a million evaluations the RSSZ value stablises but the TAE value continues to decline. Even for TAE, after approximately two million evaluations further iterations appears unnecessary, as the computational

overheads outweigh the marginal gains in fit obtained. For atypical EDs more iterations are required to allow the test statistics to stabilise. When the model switches searching from regional to national SAR, as shown in Figure 10, it takes additional time for the test statistics to catch-up with the result of using whole SAR all the way.

It appears that the RSSZ statistics of an atypical a typical ED will never converge to the same level. It is even possible that some of the few remaining discrepancies are inevitable because of the inaccuracy of the observed counts in SAS tables. Therefore it is not appropriate to set a target RSSZ value as the trigger for the termination of combinatorial optimisation; nor is it appropriate to fix the number of iterations as some EDs take far longer to 'fit' than others. Instead, a set of stopping rules have been designed to control the number of iterations. On the basis of observed test ED performance, minimum and maximum numbers of evaluations for each ED have been set at 2 million and 4.5 million respectively, giving run times on the PC used for this study of between 53 and 118 seconds per ED. Starting from 2 million evaluations, the results are evaluated at intervals of 0.5 million evaluations. After any evaluation, if all cells are deemed to fit (i.e., the number of NFC is zero) optimisation stops. This reflects the desire to devote more computing power to an ED if any cell is not satisfied. This standard of fit is extremely stringent. Some EDs will never meet it. In such cases the iteration will stop at 4.5 million evaluations.

Another innovation designed to improve the fit of abnormal EDs is that between 4 and 4.5 million evaluations, the model will not select the possible replacement element from the whole SAR any more, but use only those households already in the selected combination as potential candidates (i.e., any swap at this stage is made between households already chosen after 4 million evaluations). The theoretical consideration underpinning this procedure is within-ED homogeneity. Households close to one another tend to have similar characteristics. An ED with an abnormal distribution is likely to comprise a group of similar households whose characteristics are far from the norm. For example, suppose a given household's characteristics are so rare that only one record in the SAR can match it, but an ED contains twenty such households. Using the combinatorial optimisation routines the household has been selected, say five times after 4 million evaluations. To allow the household combination to include twenty such records, it would require more than 16 million evaluations. If, after 4 million evaluations, we limit our search to

households in the existing household combination, the required 20 duplicates are rapidly selected as the chances of picking the target household greatly increase. The result for ED DAGF12 shows that with the above procedure the five-run average TAE and RSSZ are 265 and 0.88 respectively after 4.5 million evaluations. To achieve the same level of fit it would normally require more than 6 million evaluations (see Table 8c).

6. Evaluation and comparison of the two approaches

Following the method described in Section 3, in this section each model approach is evaluated with respect to both reliability and efficiency. Model reliability depends upon the variability of model fit between runs. To assess the degree of variability associated with the estimated data, each model has been run 100 times with a different initial sample seed, generating 100 sets of synthetic microdata for each ED in the testing area. The reliability of the model's outputs are assessed though the test of goodness-of-fit between the constraints and synthetic data. Using the test statistics at cellular, tabular and general level, the fit of two sets of data generated by Pop91SR and Pop91CO can be examined at both ED (section 6.1) and ward (section 6.2) levels. Finally, the time and resource requirements of each approach can be compared to establish relative model effectiveness (section 6.3).

6.1 Comparison of outputs at ED level

For the purposes of comparison, both synthetic reconstruction and combinatorial optimisation have used nine SAS tables as constraints (listed in Table 7). The synthetic data created by Pop91SR only include the variables listed in Table 3, but the data generated by Pop91CO contain a full set of attributes within the SAR. Of the nine constraining tables, S06 and S09 cannot be recreated with the dataset generated by Pop91SR because they contain ethnic group for all residents and Table 3 only includes the ethnic group of household heads. These tables have been used as quasi-constraints in Pop91SR. Therefore the seven remaining SAS tables are used as the basis to assess the fit of synthetic data. They are tables S08, S39, S34, S35, S49, S86, S42. The counts in S08 are for all residents and they are scaled down to include residents in households only. The counts in S35 for children (aged between 0-15) are not used since Pop91SR has not included this population. For S42 only the distribution of tenure is used. In total the seven tables contain 415 cells.

First we examine whether the synthetic datasets generated by each model fit the constraints. The definition of fit has been given in Section 3.2. Table 9 shows the test results over 86 EDs in the two test wards. It reports the summary figures of the test results, giving the average numbers of NFT (non-fitting tables), NFC (non-fitting cells)

Table 9 Performance of synthetic reconstruction and combinatorial optimisation (NFT, NFC and PFC statistics)

Cookridge ward

EDCODE	Number of NFT SR	Number of NFT CO	Number of NFC SR	Number of NFC CO	Number of PFC SR	Number of PFC CO
DAFJ01	0	0	2.58	0	0.07	0
DAFJ02	0	0	3.25	0	0.14	0
DAFJ03	0	0	2.57	0	0.14	0
DAFJ04	0.01	0	2.95	0.09	0.09	0
DAFJ05	0	0	4.62	0.94	1.3	0.06
DAFJ06	0	0	4.02	0.01	0.19	0
DAFJ07	0	0	2.61	0.27	0.19	0
DAFJ08	0	0	3.77	0	0.12	0
DAFJ09	0.01	0	4.42	0.02	0.54	0
DAFJ10	0	0	3.98	0.02	0.24	0
DAFJ11	0	0	3.29	0	0.41	0
DAFJ12	0.01	0	3.89	0.01	0.2	0
DAFJ13	0	0	5.5	0.02	0.43	0
DAFJ14	0.02	0	3.63	0	0.25	0
DAFJ15	0	0	5.46	0	0.46	0
DAFJ16	0	0	4.24	0	0.25	0
DAFJ17	0	0	4.22	0.02	0.33	0
DAFJ18	0	0	3.22	0.01	0.17	0
DAFJ19	0	0	2.47	0.01	0.23	0
DAFJ20	0.01	0	3.51	0	0.17	0
DAFJ21	0	0	3.58	0	0.19	0
DAFJ22	0.02	0	4.5	0.21	0.3	0
DAFJ23	0	0	2.9	0.04	0.31	0
DAFJ24	0	0	3.15	0.04	0.47	0
DAFJ25	0.01	0	4.41	0	0.29	0
DAFJ26	0	0	5.09	0	0.27	0
DAFJ27	0	0	4.67	0	0.18	0
DAFJ28	0	0	3.53	0	0.16	0
DAFJ29	0	0	4.27	0	0.28	0
DAFJ30	0.01	0	4.65	0	0.25	0
DAFJ31	0	0	2.78	0	0.15	0
DAFJ32	0.02	0	4.11	0	0.32	0
DAFJ33	0	0	4.76	0	0.28	0
DAFJ34	0	0	5.45	0	0.57	0
DAFJ35	0.01	0	4.06	0	0.18	0
DAFJ36	0.01	0	3.36	0	0.33	0
DAFJ37	0	0	3.52	0	0.18	0
DAFJ38	0	0	4.78	0	0.5	0
DAFJ39	0	0	4.93	0.09	0.4	0

University ward

EDCODE	Number of NFT SR	Number of NFT CO	Number of NFC SR	Number of NFC CO	Number of PFC SR	Number of PFC CO
DAGF01	0.01	0	5.21	0.01	0.44	0
DAGF02	0	0	6.13	0.01	0.65	0
DAGF03	0.02	0	8.44	0.55	1.36	0.05
DAGF04	0.02	0	7.15	0.15	0.62	0
DAGF05	0.01	0	7.65	0	0.75	0
DAGF06	0	0	6.71	0	0.74	0
DAGF07	0.01	0	7.36	0.01	1.42	0
DAGF08	0.01	0	6.2	0.04	0.51	0
DAGF09	0.01	0	6.09	0.5	0.61	0.02
DAGF10	0.01	0	6.06	0	0.4	0
DAGF11	0.01	0.01	4.61	0.16	0.44	0.01
DAGF12	0	0.08	6.7	0.76	0.75	0.08
DAGF13	0	0	5.88	0.41	0.84	0
DAGF14	0.03	0	6.95	0.06	0.71	0
DAGF15	0.01	0	6.28	0.04	0.68	0
DAGF16	0.01	0	6.08	0	0.85	0
DAGF17	0.01	0	5.23	0.21	0.2	0
DAGF18	0	0	5.85	0	0.63	0
DAGF19	0.01	0	5.57	0.11	0.28	0
DAGF20	0.01	0.01	6.03	0.43	0.62	0.04
DAGF21	0	0	4.09	0.27	0.28	0
DAGF22	0	0	6.84	0.22	0.47	0
DAGF23	0.02	0	5.68	0.01	0.45	0
DAGF24	0.01	0	5.99	0.2	0.47	0
DAGF25	0.01	0	5.65	0	0.58	0
DAGF26	0.02	0	5.72	0.19	0.49	0
DAGF27	0.01	0	6.6	0.11	0.69	0
DAGF28	0.01	0	5.68	0.13	0.66	0
DAGF29	0	0	5.09	0	0.44	0
DAGF30	0.01	0	3.86	0	0.21	0
DAGF31	0.02	0	5.65	0.43	0.63	0
DAGF32	0.01	0	6.27	0.29	0.48	0
DAGF33	0.01	0	4.54	0	0.21	0
DAGF34	0.02	0	5.7	0.01	0.58	0
DAGF35	0	0	5.11	0.03	0.26	0
DAGF36	0	0	3.72	0	0.22	0
DAGF37	0.01	0	6.3	0.04	0.74	0
DAGF38	0.02	0	5.77	0.08	0.71	0
DAGF39	0.03	0	4.63	0.45	0.45	0.04
DAGF40	0.02	0	6.98	1.02	0.53	0.98
DAGF41	0.02	0	7.05	0.01	0.77	0
DAGF42	0.03	0	6.18	0.12	0.6	0
DAGF43	0.01	0	9.18	0.03	1.05	0
DAGF44	0.01	0	6.69	0.16	0.59	0
DAGF45	0	0	5.55	0.08	0.6	0
DAGF57	0	0	0.78	0.29	0.19	0.01
DAGF58	0	0.29	1.27	1.73	0.06	0.22

NFT - Not fit table
NFC - Not fit cell
PFC - Pool fit cell
SR - Synthetic reconstruction
CO - Combinatorial optimisation

Figures are mean values over 100 replications across 7 tables, comprising 415 cell counts

and PFC (poorly-fitting cells) across the 7 tables over 100 replications. The difference in performance between the two models is highlighted in Figure 11. Using SSZ as the test statistic of tabular fit it is found that both datasets fit the constraining tables extremely well. As shown in Figure 11a, for nearly half of all EDs in the test area Pop91SR produces synthetic data with the mean number of NFT equal to zero (i.e., the synthetic data fit all constraining tables for all 100 trials). The figures for the rest of the EDs are less than 0.03; a figure of 0.03 suggests that generated datasets fail to fit one table (out of 7) in three trials (out of 100). The tabular fit for the dataset generated by Pop91CO is even better. For all but four EDs the NFT values are zero; for two of the remaining four EDs the 100-run average NFT is less than 0.01. Only in two EDs (DAGF12 and DAGF58) is the fit produced by combinatorial optimisation less good than that for synthetic reconstruction.

At cellular level the number of NFC for datasets generated by Pop91SR ranges from 0.78 to 9.18 with the mean of 4.95 across 86 EDs (Figure 11b). This result means that, on average, only 5 out of a possible 415 cells fail the Z-statistic test in a given trial. However, the constraining SAS tables are modified by data blurring. If we allow for the the ±1 uncertainty over actual cell counts, the number of cells that fail Z-statistic test falls significantly. As shown in Table 9 and Figure 11c, the PFC figures are almost one tenth of the NFC numbers for the data generated by Pop91SR. On average the number of PFC is only 0.45, i.e., less than one cell poorly fit per trial. The datasets generated by Pop91CO, however, show an even better fit at individual level (see Figures 11b,c). The average numbers of NFC and PFC over 86 EDs are only 0.13 and 0.02 respectively, considerably less then that arising from synthetic reconstruction.

Having tested the goodness-of-fit of the two approaches at both tabular and cellular levels, we can now compare their overall fit in more summary form. Table 10 reports the average errors and the percentage differences of the two approaches for every test ED, measured by the mean overall RSSZ and overall TAE statistics. Both statistics show that for all but two EDs Pop91CO produce much better estimates. The average errors of the dataset generated by Pop91CO are considerably less than that of dataset created by Pop91SR. The average reductions of the overall RSSZ statistic are 81% and 66% for EDs in the Cookridge and University wards respectively. The average reductions of overall TAE are 65% and 51% for the same two wards.

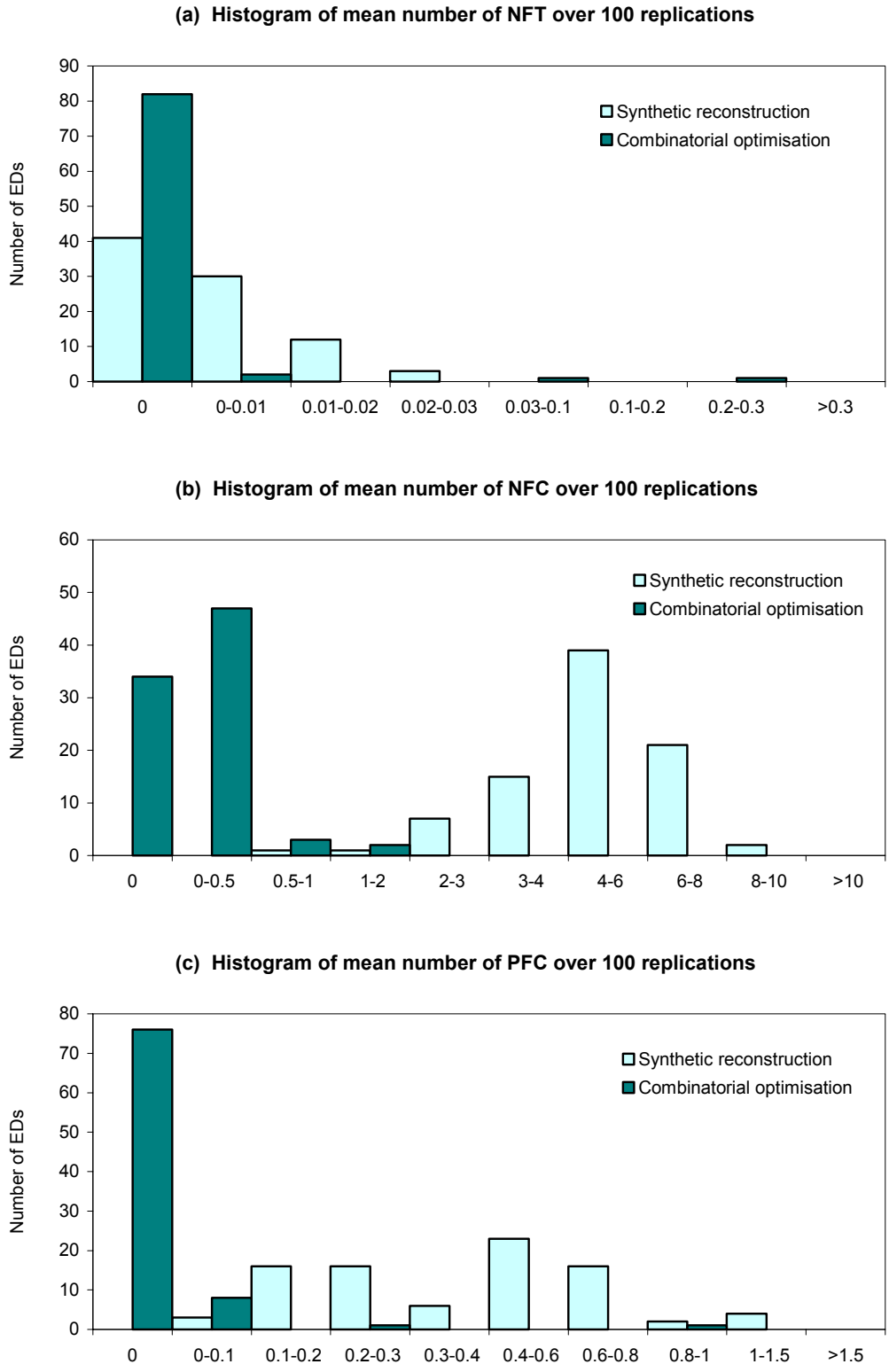


Figure 11 Performance comparison: synthetic reconstruction vs. combinatorial optimisation

Table 10 Performance of synthetic reconstruction and combinatorial optimization (RSSZ and TAE statistics)

Cookridge ward							University ward						
	RSSZ			TAE				RSSZ			TAE		
EDCODE	SR	CO	Diff. (%)	SR	CO	Diff. (%)	EDCODE	SR	CO	Diff. (%)	SR	CO	Diff. (%)
DAFJ01	1.12	0.34	*-70*	193.3	69.4	*-64*	DAGF01	1.65	0.29	*-82*	235.3	99.2	*-58*
DAFJ02	0.93	0.13	*-86*	180.1	58.0	*-68*	DAGF02	1.62	0.34	*-79*	230.1	103.8	*-55*
DAFJ03	0.70	0.14	*-80*	159.8	52.1	*-67*	DAGF03	2.05	0.42	*-80*	245.9	111.4	*-55*
DAFJ04	1.10	0.29	*-74*	184.4	69.5	*-62*	DAGF04	1.79	0.59	*-67*	227.6	125.2	*-45*
DAFJ05	1.18	0.30	*-75*	217.0	112.7	*-48*	DAGF05	1.80	0.40	*-78*	219.3	100.1	*-54*
DAFJ06	1.02	0.22	*-78*	187.1	68.0	*-64*	DAGF06	1.73	0.22	*-87*	238.4	74.6	*-69*
DAFJ07	0.95	0.33	*-65*	165.7	55.5	*-67*	DAGF07	1.93	0.30	*-84*	272.8	112.8	*-59*
DAFJ08	1.00	0.17	*-83*	170.7	59.3	*-65*	DAGF08	1.69	0.40	*-76*	235.7	80.4	*-66*
DAFJ09	1.15	0.21	*-82*	197.9	63.0	*-68*	DAGF09	1.79	0.41	*-77*	221.4	93.9	*-58*
DAFJ10	1.13	0.29	*-74*	191.8	69.6	*-64*	DAGF10	1.77	0.36	*-80*	225.6	99.2	*-56*
DAFJ11	1.00	0.10	*-90*	182.7	61.0	*-67*	DAGF11	1.31	0.55	*-58*	135.7	78.1	*-42*
DAFJ12	1.37	0.20	*-85*	204.3	59.4	*-71*	DAGF12	1.76	0.93	*-47*	248.5	160.2	*-36*
DAFJ13	1.39	0.22	*-84*	237.1	72.7	*-69*	DAGF13	1.79	0.70	*-61*	243.5	124.3	*-49*
DAFJ14	1.06	0.20	*-81*	191.6	87.9	*-54*	DAGF14	2.00	0.38	*-81*	240.9	96.7	*-60*
DAFJ15	1.49	0.16	*-89*	235.5	67.9	*-71*	DAGF15	1.78	0.25	*-86*	268.5	89.6	*-67*
DAFJ16	1.19	0.17	*-86*	223.2	82.3	*-63*	DAGF16	1.83	0.26	*-86*	275.5	89.4	*-68*
DAFJ17	1.39	0.39	*-72*	193.8	95.8	*-51*	DAGF17	1.54	0.42	*-73*	184.1	75.2	*-59*
DAFJ18	1.12	0.24	*-79*	205.5	79.6	*-61*	DAGF18	1.61	0.26	*-84*	240.0	83.7	*-65*
DAFJ19	0.77	0.17	*-78*	172.4	64.8	*-62*	DAGF19	1.52	0.51	*-66*	188.3	94.6	*-50*
DAFJ20	1.06	0.17	*-84*	189.4	64.2	*-66*	DAGF20	1.45	0.54	*-63*	187.5	99.5	*-47*
DAFJ21	1.06	0.16	*-85*	206.1	66.4	*-68*	DAGF21	1.51	0.67	*-56*	176.6	93.0	*-47*
DAFJ22	1.48	0.25	*-83*	226.7	65.6	*-71*	DAGF22	1.61	0.55	*-66*	217.4	116.0	*-47*
DAFJ23	0.87	0.29	*-67*	181.1	103.6	*-43*	DAGF23	1.66	0.35	*-79*	215.9	87.2	*-60*
DAFJ24	1.12	0.22	*-80*	183.2	63.0	*-66*	DAGF24	1.73	0.58	*-66*	170.6	101.8	*-40*
DAFJ25	1.32	0.16	*-88*	216.0	69.1	*-68*	DAGF25	1.88	0.36	*-81*	273.6	120.6	*-56*
DAFJ26	1.34	0.23	*-83*	203.3	66.4	*-67*	DAGF26	1.79	0.34	*-81*	229.0	94.0	*-59*
DAFJ27	1.30	0.26	*-80*	209.0	79.3	*-62*	DAGF27	1.97	0.35	*-82*	264.9	102.0	*-61*
DAFJ28	1.16	0.23	*-80*	195.9	75.1	*-62*	DAGF28	1.77	0.36	*-80*	239.3	100.3	*-58*
DAFJ29	1.19	0.25	*-79*	215.7	71.1	*-67*	DAGF29	1.62	0.35	*-78*	205.8	78.8	*-62*
DAFJ30	1.15	0.18	*-84*	212.0	64.3	*-70*	DAGF30	1.12	0.29	*-74*	156.9	63.8	*-59*
DAFJ31	0.85	0.19	*-78*	172.1	59.1	*-66*	DAGF31	1.56	0.46	*-71*	182.9	76.5	*-58*
DAFJ32	1.25	0.22	*-82*	191.8	75.4	*-61*	DAGF32	1.66	0.39	*-77*	218.8	70.6	*-68*
DAFJ33	1.34	0.18	*-87*	225.8	68.9	*-69*	DAGF33	1.39	0.17	*-88*	205.1	63.4	*-69*
DAFJ34	1.30	0.26	*-80*	230.4	79.6	*-65*	DAGF34	1.54	0.32	*-79*	216.4	74.0	*-66*
DAFJ35	1.21	0.19	*-84*	205.7	72.2	*-65*	DAGF35	1.54	0.28	*-82*	196.2	72.9	*-63*
DAFJ36	1.07	0.17	*-84*	179.9	64.1	*-64*	DAGF36	1.20	0.28	*-77*	169.7	72.2	*-57*
DAFJ37	1.12	0.20	*-82*	203.1	57.5	*-72*	DAGF37	1.88	0.28	*-85*	279.5	106.1	*-62*
DAFJ38	1.22	0.11	*-91*	216.9	57.6	*-73*	DAGF38	1.76	0.33	*-81*	242.4	120.6	*-50*
DAFJ39	1.34	0.19	*-86*	239.2	61.3	*-74*	DAGF39	1.62	0.56	*-65*	185.6	103.5	*-44*
							DAGF40	1.56	0.48	*-69*	206.0	113.4	*-45*
Average			*-81*			*-65*	DAGF41	1.94	0.32	*-84*	293.3	116.7	*-60*
							DAGF42	1.88	0.44	*-77*	226.6	110.4	*-51*
							DAGF43	1.90	0.36	*-81*	271.0	107.3	*-60*
Figures are mean values over 100 replications							DAGF44	1.85	0.49	*-74*	237.7	135.3	*-43*
Total number of tables: 7							DAGF45	1.63	0.69	*-58*	222.5	177.0	*-20*
Total number of cells: 415							DAGF57	0.20	0.38	*90*	25.1	38.3	*53*
SR - Synthetic reconstruction							DAGF58	0.48	1.35	*181*	64.3	100.7	*57*
CO - Combinatorial optimisation													
							Average			*-66*			*-51*

The two EDs that Pop91CO fail to produce better estimates than Pop91SR are the student EDs DAGF57 and DAGF58, in the aptly named University ward. These two EDs are extremely atypical; their distance from norm have been identified as the second and third highest in England and Wales (Voas and Williamson, 2001b). ED DAGF57 has only 20 households but comprises approximate 171 rooms in total. About 331 students were in this ED on the census night, but only 54 of them are classified as household residents, the rest being visitors (non-residents). ED DAGF58 has similar characteristics, comprising 46 households and 135 residents in households. The reasons that the estimates of Pop91CO are not as good as that of Pop91SR for these highly atypical EDs are twofold. First, the required households might be so unusual that there are no records in the SAR that can match them. In that case Pop91CO selects the household combination that produces the least discrepancy between the estimated and observed data within the time allowed. The size of the discrepancy largely depends on the whether or not a closest match exists. The test statistics for ED DAGF57 are actually very good, the mean RSSZ value is 0.38, slightly higher than the average figure (0.33) over the test area. In contrast this figure is 1.35 for ED DAGF58, the highest in the test area, indicating that there are less households available in the SAR that match the ED's characteristics. Nevertheless, this statistic is still lower than the average RSSZ (1.41) for the test area produced by Pop91SR.

Second, Pop91SR actually produces its best estimates for these two EDs. The mean RSSZ statistics are only 0.2 and 0.48 for ED DAGF57 and ED DAGF58 respectively. In Pop91SR, the generation of a variable depends on the conditional probability distributions and the random number. When a conditional distribution is so uneven that only one category is one and the others are zero, the assignment of the new variable will not be affected by the random number since there is only one choice. EDs DAGF57 and DAGF58 consist of a small set of highly similar households. Hence the constraining tables contain many empty cells, reducing the number of possible choices to sample from, so reducing the impact of random sampling error.

The above tests were performed for each synthetic dataset generated by a different initial sample seed, and summarised giving the mean fit over 100 replications. In constrast, Table 11 reports for each approach the overall RSSZ of their 100-run means, thereby giving the fit of the mean, rather than the mean fit. These figures are all very small and

Table 11 Performance of synthetic reconstruction and combinatorial optimization (RSSZ of mean)

Cookridge ward				University ward			
	RSSZ of mean				RSSZ of mean		
EDCODE	SR	CO	CO-SR	EDCODE	SR	CO	CO-SR
DAFJ01	0.21	0.21	*0.00*	DAGF01	0.14	0.16	*0.01*
DAFJ02	0.10	0.06	*-0.04*	DAGF02	0.16	0.15	*-0.01*
DAFJ03	0.05	0.07	*0.02*	DAGF03	0.29	0.22	*-0.06*
DAFJ04	0.13	0.16	*0.03*	DAGF04	0.20	0.28	*0.09*
DAFJ05	0.24	0.20	*-0.04*	DAGF05	0.22	0.21	*-0.01*
DAFJ06	0.12	0.12	*0.01*	DAGF06	0.13	0.08	*-0.05*
DAFJ07	0.12	0.19	*0.06*	DAGF07	0.24	0.14	*-0.10*
DAFJ08	0.10	0.08	*-0.02*	DAGF08	0.17	0.22	*0.05*
DAFJ09	0.10	0.10	*0.00*	DAGF09	0.22	0.23	*0.01*
DAFJ10	0.15	0.15	*0.00*	DAGF10	0.16	0.15	*-0.01*
DAFJ11	0.09	0.05	*-0.03*	DAGF11	0.22	0.24	*0.02*
DAFJ12	0.15	0.08	*-0.08*	DAGF12	0.20	0.59	*0.38*
DAFJ13	0.14	0.09	*-0.05*	DAGF13	0.15	0.42	*0.27*
DAFJ14	0.07	0.10	*0.03*	DAGF14	0.15	0.16	*0.01*
DAFJ15	0.23	0.08	*-0.16*	DAGF15	0.14	0.11	*-0.03*
DAFJ16	0.20	0.10	*-0.10*	DAGF16	0.17	0.13	*-0.03*
DAFJ17	0.17	0.19	*0.02*	DAGF17	0.16	0.19	*0.03*
DAFJ18	0.12	0.10	*-0.02*	DAGF18	0.14	0.09	*-0.05*
DAFJ19	0.09	0.09	*-0.01*	DAGF19	0.13	0.22	*0.09*
DAFJ20	0.09	0.07	*-0.02*	DAGF20	0.20	0.28	*0.08*
DAFJ21	0.17	0.08	*-0.10*	DAGF21	0.21	0.38	*0.17*
DAFJ22	0.10	0.12	*0.02*	DAGF22	0.13	0.29	*0.15*
DAFJ23	0.11	0.16	*0.06*	DAGF23	0.14	0.19	*0.04*
DAFJ24	0.11	0.12	*0.01*	DAGF24	0.22	0.25	*0.03*
DAFJ25	0.09	0.08	*-0.01*	DAGF25	0.13	0.16	*0.02*
DAFJ26	0.12	0.09	*-0.03*	DAGF26	0.12	0.14	*0.02*
DAFJ27	0.19	0.14	*-0.06*	DAGF27	0.15	0.16	*0.01*
DAFJ28	0.12	0.11	*-0.01*	DAGF28	0.12	0.17	*0.05*
DAFJ29	0.18	0.15	*-0.03*	DAGF29	0.21	0.19	*-0.02*
DAFJ30	0.10	0.09	*-0.01*	DAGF30	0.14	0.14	*0.00*
DAFJ31	0.08	0.09	*0.00*	DAGF31	0.28	0.32	*0.04*
DAFJ32	0.15	0.13	*-0.03*	DAGF32	0.23	0.16	*-0.07*
DAFJ33	0.16	0.09	*-0.07*	DAGF33	0.17	0.11	*-0.06*
DAFJ34	0.15	0.13	*-0.01*	DAGF34	0.17	0.13	*-0.04*
DAFJ35	0.16	0.10	*-0.07*	DAGF35	0.21	0.15	*-0.05*
DAFJ36	0.10	0.10	*-0.01*	DAGF36	0.15	0.22	*0.07*
DAFJ37	0.11	0.10	*-0.01*	DAGF37	0.14	0.11	*-0.03*
DAFJ38	0.11	0.05	*-0.07*	DAGF38	0.14	0.15	*0.01*
DAFJ39	0.15	0.09	*-0.06*	DAGF39	0.15	0.29	*0.14*
				DAGF40	0.22	0.32	*0.09*
				DAGF41	0.18	0.17	*-0.01*
Number of replications: 100				DAGF42	0.17	0.20	*0.04*
Total number of tables: 7				DAGF43	0.31	0.20	*-0.12*
Total number of cells: 415				DAGF44	0.19	0.29	*0.10*
SR - Synthetic reconstruction				DAGF45	0.28	0.43	*0.15*
CO - Combinatorial optimisation				DAGF57	0.12	0.17	*0.05*
				DAGF58	0.07	0.98	*0.91*

both models appear to produce similar levels of fit to the expected values. For the dataset generated by Pop91SR the average values of the RSSZ of mean are 0.13 and 0.18 for Cookridge and University ward respectively. For the dataset generated by Pop91CO equivalent figures are 0.11 and 0.23. The greater gap in RSSZ between wards indicates that the output of Pop91CO is slightly more sensitive to location than that of Pop91SR.

Our analysis suggests that both models can produce nearly unbiased estimates (i.e. the expected synthetic counts are equal or very close to the observed counts), but the variances are very different. The level of variance of the synthetic data generated by Pop91CO is considerably lower than that of the microdata created by Pop91SR. With hindsight, this finding can be readily explained. The variance in the data generated by Pop91SR mainly arises from random sampling. The household combination generated by Pop91CO, however, is the final result of millions of iterations. Although altering the sample seeds would produce a different final household combination, the effect on random sampling is not as great.

Table 12 presents an example of the fit achieved by two models to SAS table 34 for ED DAFJ01 in the Cookridge ward. Both models produce excellent results in terms of mean estimates. The SSZ of the 100-run means show that two models present the same level of fit on the expected counts. But tests on individual datasets reveal the difference in the degree of variability between the two sets of data. Z statistic tests show that for the synthetic data generated by Pop91SR 31 out of 40 cells never fail the test; on average less than one cell produces data with a Z score exceeding the 5% critical values. But the data generated by Pop91CO fit all 40 cells in all trials. Further, we can see the difference from the estimated 95% confidence interval for each count (i.e., the range within which 95% of synthetic values lie). As shown in Table 12a, the confidence interval for the data generated by Pop91CO is always less than that of the data generated by Pop91SR. Only in two out of forty cells is the width of the 95% confidence interval for Pop91CO three, the remaining cells having intervals of two or less. This suggests that the synthetic data generated by Pop91CO are more reliable since we can locate their usual value within very narrow bands. At the tabular level the mean SSZ for the data generated by Pop91SR is 12.8, which is considerably less than 55.8, the critical value. But once again Pop91CO performs better, producing a mean SSZ of only 2.4.

Table 12 Comparing the fit of estimated population for ED DAFJ01 to SAS table 34

(a) Cellular test		Synthetic reconstruction				Combinatorial optimisation			
	SAS Table 34	Mean synthetic	Top & bottom of 95% interval		% of |Z|>1.96	Mean synthetic	Top & bottom of 95% interval		% of |Z|>1.96
Male, single, widowed, divorced									
Employees-full time	17	18.8	23	14	0	18.0	19	17	0
Employees-part time	4	4.2	8	2	3	4.0	4	4	0
Self emp.-with employees	0	0.0	0	0	0	0.0	0	0	0
Self emp.-without employees	4	4.5	7	3	2	4.0	5	4	0
On a government scheme	1	0.0	0	0	0	1.0	1	1	0
Unemployed	1	1.0	3	0	5	1.1	2	1	0
Students	8	8.3	12	5	0	8.0	8	7	0
Permanently sick	1	1.1	3	0	8	1.0	1	1	0
Retired	12	12.9	15	11	0	12.0	13	11	0
Other inactive	1	0.9	2	0	1	1.0	1	1	0
Male, married									
Employees-full time	63	63.4	68	60	0	62.3	64	61	0
Employees-part time	10	10.1	14	8	0	9.8	10	9	0
Self emp.-with employees	11	10.9	15	7	0	9.7	10	9	0
Self emp.-without employees	6	6.1	9	3	0	6.2	7	6	0
On a government scheme	0	0.0	0	0	0	0.0	0	0	0
Unemployed	1	1.0	3	0	7	1.7	2	1	0
Students	1	0.7	2	0	1	1.0	1	1	0
Permanently sick	6	5.5	8	3	0	6.0	7	5	0
Retired	41	41.0	44	38	0	41.1	42	40	0
Other inactive	0	0.0	0	0	0	0.0	0	0	0
Female, single, widowed, divorced									
Employees-full time	21	21.5	26	18	0	20.6	21	20	0
Employees-part time	10	10.6	15	6	0	9.8	11	9	0
Self emp.-with employees	0	0.0	0	0	0	0.0	0	0	0
Self emp.-without employees	0	0.0	0	0	0	0.0	0	0	0
On a government scheme	0	0.0	0	0	0	0.0	0	0	0
Unemployed	0	0.0	0	0	0	0.0	0	0	0
Students	10	10.5	13	8	0	10.7	11	10	0
Permanently sick	1	1.1	3	0	4	1.0	1	1	0
Retired	17	18.1	21	16	0	17.7	18	17	0
Other inactive	9	9.5	13	6	2	9.3	10	9	0
Female, married									
Employees-full time	23	22.3	26	19	0	22.9	24	22	0
Employees-part time	39	38.3	43	33	0	38.5	39	37	0
Self emp.-with employees	2	2.0	4	0	0	1.4	2	1	0
Self emp.-without employees	7	7.0	10	4	0	6.2	7	6	0
On a government scheme	0	0.0	0	0	0	0.0	0	0	0
Unemployed	2	1.7	4	1	0	1.0	1	1	0
Students	0	0.0	0	0	0	0.0	0	0	0
Permanently sick	3	3.0	6	1	0	3.0	4	3	0
Retired	32	31.2	34	28	0	31.5	33	30	0
Other inactive	34	32.9	38	29	0	33.7	35	33	0

(b) Tabular test	Synthetic reconstruction	Combinatorial optimisation
SSZ of mean	1.8	1.6
Mean TAE	37.2	11.9
Mean SSZ	12.8	2.4
% of SSZ > Critical value*	0	0
Mean NFC	0.3	0
Mean PFC	0	0

* 5% chi-square critical value = 55.8; Number of replications = 100,

6.2 Comparison of outputs at ward level

The second test for the synthetic populations is how well they fit the constraints at ward level. In other words, is the fit degraded (or improved) if the synthetic data are aggregated to ward level? Interestingly, this element of model fit has never been examined before. We also wish to compare the two approaches' performance at ward level in capturing the interaction between constraining variables.

First, we examine the fit of both Pop91C and Pop91SR results to those tables that are available at both ward and ED level (i.e., excluding table L45). As suggested in Section 3.2, the observed ward-level counts are the taken to be aggregations of the corresponding SAS table counts within the ward, to avoid any data inconsistency between LBS table counts and aggregated SAS table counts due to data blurring. Therefore, the synthetic data are aggregated to the SAS table format rather than the LBS format. Table 13 reports the test results for the two sets of synthetic data. The summary statistics over seven tables show that Pop91CO produces much more accurate estimates than Pop91SR. The dataset generated by Pop91CO can fit not only all the tables but also almost every cell as well – on average only two cells out of 415 per replication are 'non-fitting' cells. Given the level of fit at ED level it is perhaps not surprising that at ward level the degree of variation of Pop91CO's output is considerably less than that of Pop91SR.

Overall figures also show that, in general, at ward level the mean distribution of the 100 synthetic reconstruction estimates is closer to the target distribution than that for combinatorial optimisation (RSSZ of mean). However, in almost every other respect combinatorial optimisation offers superior performance, in particular offering markedly reduced variance (lower average numbers of non-fitting cells and tables). The same story is true if fit is analysed at tabular level (Table 13b). The extreme case is table S35. For this table datasets generated by Pop91SR perform particularly poorly. In University ward, even though the SSZ of mean is only 3.8 for Pop91SR, less than half of the value obtained using Pop91CO, out of 100 synthetic reconstruction trials 74% fail to fit ward-level table 35 in University ward, compared to 0% for combinatorial optimisation.

For most purposes only a single set of synthetic microdata will be used. Therefore a guaranteed close fit (minimal variability) is to much to be preferred to assurances of

Table 13 Performance of synthetic reconstruction and combinatorial optimisation at ward level

(a) Overall fit

	Cookridge ward					University ward				
	Overall TAE	Overall RSSZ	Number of NFT	Number of NFC	RSSZ of mean	Overall TAE	Overall RSSZ	Number of NFT	Number of NFC	RSSZ of mean
Synthetic reconstruction	2307	2.98	0.17	15.6	0.44	2701	3.64	0.8	19.3	0.31
Combinatorial optimisation	1084	0.84	0	2.3	0.64	1498	1.28	0	1.9	1.05

(b) Tabular fit

			Cookridge ward					University ward				
Table	Number of cells	Critical value	TAE	SSZ and % of SSZ > critical value		Number of NFC	SSZ of mean	TAE	SSZ and % of SSZ > critical value		Number of NFC	SSZ of mean
Synthetic reconstruction												
39	*28*	*41.3*	6	0	0	0	0	6	0.1	0	0.0	0.1
35	*68*	*88.3*	687	67.5	14	4.3	3.3	836	108.4	74	7.7	3.8
34	*40*	*55.8*	401	29.6	3	1.1	2.8	399	28.1	2	0.9	1.8
8	*180*	*212.3*	768	141.1	0	8.4	40.2	835	161.7	4	8.9	40.6
42	*7*	*14.1*	85	3.0	0	0.1	0.2	112	4.3	0	0.1	0.1
49	*16*	*26.3*	99	12.7	0	0.5	3.7	164	11.7	0	0.3	0.8
86	*76*	*97.4*	262	31.8	0	1.2	1.4	348	38.1	0	1.4	0.9
Combinatorial optimisation												
39	*28*	*41.3*	50	0.6	0	0	0.4	68	1.0	0	0	0.6
35	*68*	*88.3*	240	7.8	0	0	5.5	273	12.3	0	0.0	8.9
34	*40*	*55.8*	197	4.8	0	0	5.3	260	12.9	0	0.2	14
8	*180*	*212.3*	393	60.3	0	1.7	39.1	505	86.4	0	1.6	60.2
42	*7*	*14.1*	40	0.9	0	0	0.7	100	3.2	0	0	2.9
49	*16*	*26.3*	28	2.8	0	0	2.1	75	2.3	0	0	1.8
86	*76*	*97.4*	138	18.9	0	0.7	14.9	218	15.7	0	0.1	12.1

Critical values are table-specific 5% critical values (degrees of freedom = number of cells)
Test statistics are averages over 100 replications. Number of cells in all tables = 415

minimum bias in the estimation process (best 100-run mean). The main cause of the greater variability in results obtained from synthetic reconstruction across all tables is the impact of random sampling at ED level, which is amplified when cumulated from ED to ward level. Table S35 suffers from the additional problem that, during synthetic reconstruction, its counts were constrained to agree to coarser age groupings in table S39 (to eliminate discrepancies introduced by census data blurring). The adjustment process improves fit to S39, but at the expense of some marginal decrease in fit to S35. At ED level this is unimportant, but once again becomes magnified when cumulated to ward level.

A second test is to examine the fit of the aggregated ward-level data to LBS table 45. L45 cross-tabulates four variables: age, sex and economic position of household head and tenure, from which we can derive four tables. L45a contains all four variables, L45b comprises age and sex of head and tenure, L45c includes age, sex and economic position of head, and L45d gives only tenure by economic position of head. When measuring the fit to these tables some errors may be contributed by the data inconsistency due to data blurring and the restrictions of the synthetic dataset. For example, populations in special EDs are not included in the synthetic data, so the total of L45 (the number of households in the ward) and the synthetic total may be not identical. Nevertheless, these tables allow us to examine how well the synthetic data capture the interactions between different variable combinations.

Table 14a reports the test statistics for the datasets generated by Pop91SR. Although most of the synthetic data fit the tables L45a-d, the failure-to-fit rates are higher than previously found in other tables. The numbers of synthetic data failing to fit L45a are 38% and 32% for the Cookridge and University wards respectively. The fit is slightly improved when the table is collapsed into L45b, L45c and L45d. The best fit is found in the Cookridge ward on L45b, where all the synthetic data passed the test of tabular fit. An explanation is that the housing tenure in this ward is dominated by owner occupied, representing 72% households. So the variance of allocating this variable is relatively lower, resulting in good fit on tenure.

If we don't use L45 as a constraining table in Pop91SR, then we can assess the model's ability, unaided, to reassemble the interrelationship between the variables contained in

Table 14 Fit to LBS Table 45

LBS Table	Cookridge ward TAE	SSZ and % of SSZ > critical value		Number of NFC	SSZ of mean	University ward TAE	SSZ and % of SSZ > critical value		Number of NFC	SSZ of mean
(a) Synthetic reconstruction										
L45a	526	119.6	*38*	7.5	66	663	113.9	*32*	7.2	60.7
L45b	297	31.1	*0*	1.2	22.8	418	41.3	*26*	2.6	30.1
L45c	255	33.2	*34*	2.6	19.3	292	23.2	*12*	1.1	11.3
L45d	198	18.7	*35*	1.3	10.7	278	16.6	*21*	1.1	7.7
(b) Synthetic reconstruction (without using L45 as constraint)										
L45a	478	174.6	*99*	10.5	114.1	871	261	*100*	20.7	203
L45b	220	27.6	*3*	1.2	17.8	527	95.9	*100*	7.6	83.9
L45c	258	35.1	*37*	2.9	21.1	321	32.4	*32*	2.4	20.1
L45d	210	18.8	*32*	1.4	10.5	314	29.3	*83*	2.6	20.1
(c) Combinatorial optimisation										
L45a	503	190.3	*100*	13.6	147.3	777	252.6	*100*	17.6	201.4
L45b	319	37.4	*16*	1.8	28.1	463	62.8	*93*	4.1	49.8
L45c	217	30.2	*13*	2.7	25.1	350	39.0	*65*	3.0	34.7
L45d	150	22.8	*62*	1.6	20.9	257	33.6	*98*	2.0	31.4

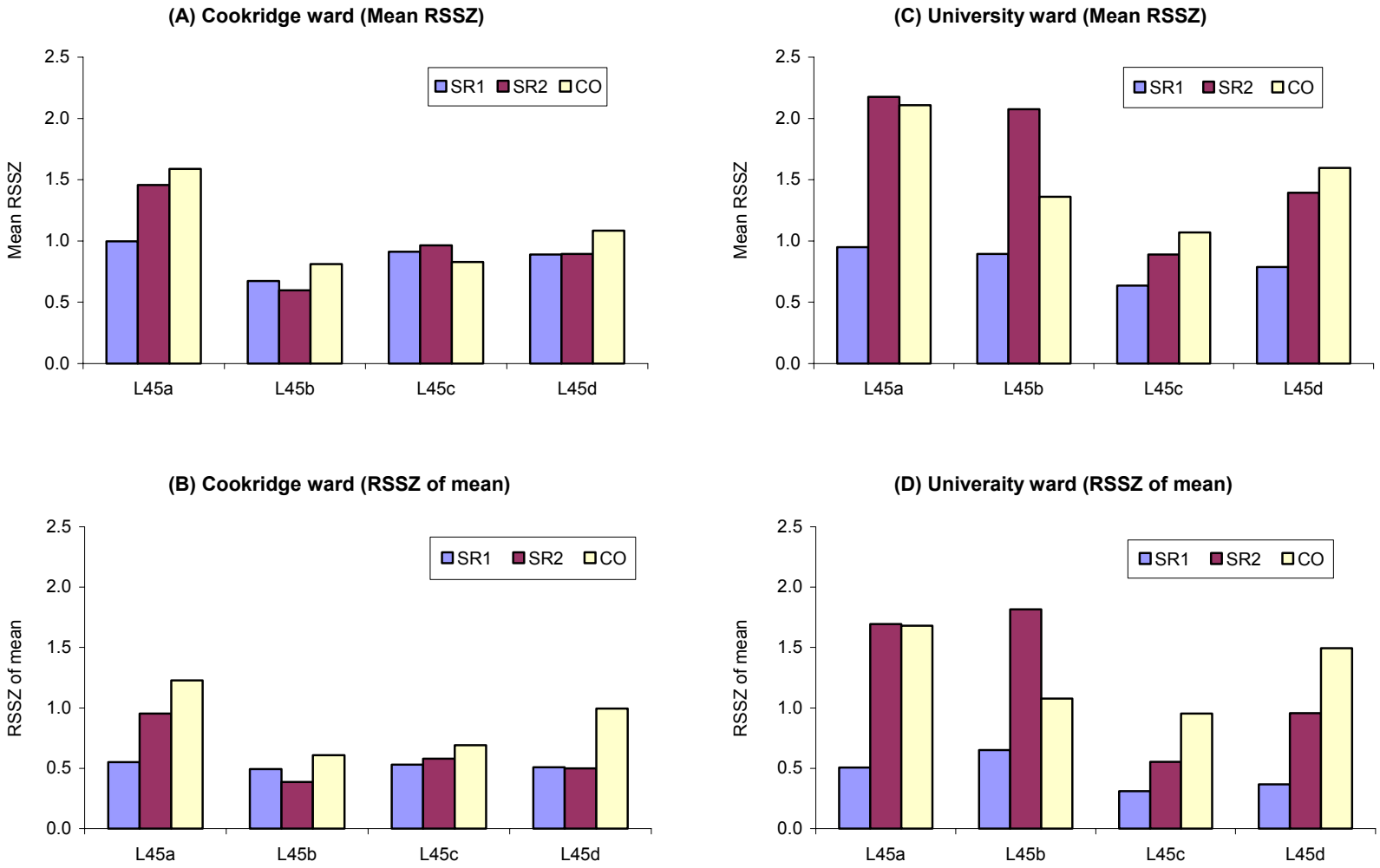
Note: The four tables are all drawn from L45. Each table's variables, number of cells and critical chi-square value for are as follows:
(age, sex, and economic position are for the heads of household)

	Variable	*Number of cells*	*Critical value*
L45a	Age/sex/economic position/tenure	96	119.9
L45b	Age/sex/tenure	32	46.2
L45c	Age/sex/economic position	24	36.4
L45d	Economic position/tenure	12	21.0

L45. Among the four variables in L45 only the relationship between age and sex of head is constrained. The univariate distribution of tenure is also a constrained variable. Economic position of head is not constrained, but economic position of household residents is. Therefore, L45b (age/sex/tenure) can be viewed as a cross-tabulation of constrained variables, whilst the remaining three include only unconstrained (L45c, L45d) or partially constrained (L45a) variables. Table 14b reports the test results based on synthetic data created by Pop91SR without L45 as constraint. It is apparent that the fit is not as good as that previously found with L45 as constraint. In particular the synthetic data fail to fit L45a in almost every trial, highlighting the difficulty of capturing the interactions of all four variables.

Table 14c reports the test results on synthetic data created by combinatorial optimisation. In general the fit is similar to the result of the dataset generated by Pop91SR without using L45. Figure 12 highlights the difference in performance between the three sets of synthetic data. This time we plot the figures of the mean RSSZ and RSSZ of mean estimates so that the average error and bias of the three datasets can be compared directly. SR1 is the data generated by Pop91SR, SR2 is the data generated by Pop91SR without L45 as constraint, and CO is the data generated by Pop91CO. The performance of SR1 is generally acceptable, as all the values of mean RSSZ are less than one and the figures of RSSZ of mean are significantly lower than one. The performances of SR2 and CO on the tabulation of constrained variables (L45b) are as good as SR1 for the Cookridge ward, but worse than SR1 for the University ward. Between SR2 and CO, the fit of SR2 to L45b is significantly poorer than that of CO. The fits of SR2 and CO on the other tables containing unconstrained variable are generally poor.

Unless the variables are highly correlated with those that were chosen, any unconstrained relationship produced by the synthetic data at lower geographical level tends to follow the distribution at higher geographical levels. Without using L45 as the constraint the fit to this table, therefore, largely depends on how far the cross-distribution of the variables involved differs from the national distribution. The greater the divergence, the less likely the synthetic data are to produce a suitable match. For this reason it is no surprise that the fit on these tables for the University ward is poorer than that for the Cookridge ward.



SR1 - Synthetic reconstruction **SR2** - Synthetic reconstruction (without using L45 as constraint) **CO** - Combinatorial optimisation

Figure 12 Comaprison of the fit to LBS table 45

Previous work by Voas and Williamson (2000a) demonstrate that synthetic data generated by combinatorial optimisation can produce very good fit on unconstrained cross-tabulations of variables involved in the original set of constraining tables at ED level. But the fit to tabulations of variables not used as constraints is generally poor, suggesting the problem of fitting tables with unconstrained or partially variables is unavoidable. Our test at ward level suggests that for a typical area, such as Cookridge ward, synthetic data generated by combinatorial optimisation can still produce good fit on unconstrained tables between constrained variables at ward level (L45b). This provides us with some confidence when using the synthetic data to estimate the unknown relationships. But for atypical areas, such as University ward, the fit on unconstrained tabulations of constrained variables may be degraded if the data are aggregated from ED to ward level geographies. On the other hand, even this negative conclusion should not be overstated, as it is partly a product of the stringent nature of our definition of fit. Table 15 allows further examination of the fit of Pop91CO output to L45b at cellular level for University ward. This reveals that on average only four out of 32 cells produce synthetic counts with Z scores exceeding the 5% critical value. The discrepancies are most apparent for female heads between 45 and personable age, living in the their own house. In view of the heavy concentration of students in the area it is unsurprising that the actual count is less than estimates. But even for this difficult to fit ward, it is perhaps reassuring to note that the general distribution of the unconstrained relationship between age/sex and tenure broadly follows the expected. Mean synthetic counts are low (~50) when expected counts are low and high (~500) when high.

6.3 Efficiency

The development of a small-area population reconstruction model takes a considerable period of time. It took four to five months for an experienced programmer to develop Pop91SR, which currently includes 'only' nine attributes. It is estimated that at least one week would be needed to add an additional variable. Each additional conditional probability has to be specifically tailored to fit available data at ward and ED levels, including counts in previously added constraining tables. The combinatorial optimisation approach, however, is easier to standardise to suit different constraining tables. There is no need to adjust the constraining tables, because the selected household combination is the one producing the smallest discrepancy between estimated and observed data within

Table 15 Fit of synthetic data generated with combinatorial optimisation to table L45b

Age of head	16 - 29				30 - 44				45 - pensionable age				pensionable age			
Tenure	Owner Occupied	Rented (Privately/ with a job)	Rented (housing assoc.)	Rented (council)	Owner Occupied	Rented (Privately/ with a job)	Rented (housing assoc.)	Rented (council)	Owner Occupied	Rented (Privately/ with a job)	Rented (housing assoc.)	Rented (council)	Owner Occupied	Rented (Privately/ with a job)	Rented (housing assoc.)	Rented (council)
(a) Male heads																
Actual count	198	637	122	339	318	321	108	506	349	152	65	687	143	84	46	598
Mean synthetic	183	628	113	343	354	297	107	482	353	149	64	698	145	76	47	576
Top of 95% interval	200	650	126	360	376	314	122	503	372	166	75	716	162	86	58	590
Bottom of 95% interval	165	611	101	327	331	279	93	462	337	132	50	678	131	64	36	557
Z of Mean	-1.07	-0.39	-0.83	0.23	**2.08**	-1.37	-0.07	-1.09	0.23	-0.23	-0.19	0.45	0.19	-0.86	0.20	-0.94
% of |Z| > 1.96	6	0	1	0	67	10	1	2	0	0	2	0	2	6	3	1
(b) Female heads																
Actual count	134	501	209	369	154	105	112	377	70	45	48	312	125	87	115	917
Mean synthetic	107	484	203	401	145	131	120	355	105	48	45	321	113	84	109	899
Topof95%interval	124	503	221	420	160	147	137	372	119	62	55	340	128	94	118	916
Bottomof95%intvl	94	466	187	385	126	118	108	338	91	39	35	297	101	73	99	883
Z of Mean	**-2.37**	-0.78	-0.44	1.72	-0.76	**2.57**	0.79	-1.19	**4.15**	0.38	-0.37	0.49	-1.05	-0.37	-0.58	-0.62
% of |Z| > 1.96	67	0	0	44	7	73	7	1	100	6	1	0	5	0	1	0

the time allowed. The time taken to develop the main Pop91CO program suite is estimated to have taken not less than that of Pop91SR, but adding another constraining table only takes one or two days. The average computing time over the test area with Pop91SR is 1.3 seconds per ED. The running time with Pop91CO is 69.6 seconds per ED, considerably higher than Pop91SR, if within acceptable limits for generating microdata for a large area.

A distinguishing characteristic of combinatorial optimisation is its flexibility of selecting the constraining tables. Synthetic reconstruction modelling is a step by step process, with data created following in a fixed order. Combinatorial optimisation does not have this restriction, so users are able to select constraining tables/variables according to their own requirements, hence producing bespoke microdata. Another strength of combinatorial optimisation, which cannot be overemphasised, is that the synthetic individuals are automatically nested into families and households (when using household SAR as the parent population). In contrast, the dataset created by Pop91SR, as described in this paper, comprises simply a list of individuals. Going a step further to generate family and household membership would be highly problematic. Assumptions, typically subjective, would have to be made about the nature of the relationships between the family members. For example, a spouse might only be assigned to a married head if they share the same ethnic origin (see Birkin and Clarke, 1988). This will certainly affect the accuracy of model output. An alternative solution is to generate the attribute of household composition given the household head attributes already captured, and then to add appropriate family members to match. The problem with this is that there is a shortage of data that link household composition with the characteristics of the household head. Only one SAS table (S43) links household composition with ethnic group of household head. Consequently, most of the relationships required for generating household composition would have to be assumed to follow national distributions. In addition, as already demonstrated, the fit of synthetic data to unconstrained relationships is not always satisfactory. The error generated at this stage would certainly affect the estimates of family structure such as spouse, dependants and non-dependants.

7. Conclusion

The work reported here offers for the first time a thorough comparison of two established methodologies, synthetic reconstruction and combinatorial optimisation, for the creation of small area synthetic microdata, presenting at the same time new developments in each approach. Two computer models, Pop91SR and Pop91CO, have been developed for the reconstruction of ED level populations drawing upon 1991 Census data. Pop91SR is a new model based upon the synthetic reconstruction method, whilst Pop91CO is the latest version of combinatorial optimisation model. Each model benefits from methodological innovations designed to improve the accuracy and consistency of the outputs.

Compared with previous synthetic reconstruction models, Pop91SR employs the following new techniques: (a) using the SAR to examine the relationship between variables and determine the ordering of conditional probabilities; (b) employing a three-level modelling approach to create the conditional distributions, combining data from the SAS, LBS and SAR; and (c) adopting a modified Monte Carlo sampling procedure. These techniques maximise the use of information and greatly reduce the sampling error, thereby increasing estimation accuracy.

The major improvements in Pop91CO are: (a) using a new criterion ($RSSZ_m$) for the selection of household combination; (b) increasing the regional representation of the selected households by using regional SAR at the first stage; and (c) designing a set of stopping rules to control the number of iterations and improve the consistency of outputs. Using $RSSZ_m$ as the selection criterion is the major breakthrough in combinatorial optimisation modelling, yielding significant improvements in the quality of the synthetic data generated.

An assessment of outputs from the two rival approaches, produced using the same small area constraints, suggests that both can produce synthetic microdata that fit constraining tables extremely well. But further examination of the dispersion of the synthetic data has shown that the variability of datasets generated by combinatorial optimisation is considerably less than that for datasets created by synthetic reconstruction, at both ED and ward levels. The fundamental problem for the synthetic reconstruction approach is that it is a Monte Carlo based approach subject to sampling error. In contrast, the outputs

of separate combinatorial optimisation runs, arising from an intelligent search heuristic, are much less variable and, hence, individually much more reliable. In addition, combinatorial optimisation permits much greater flexibility in selecting small area constraints. Perhaps of even greater importance for many uses, combinatorial optimisation automatically places synthetic individuals within families and households whilst reflecting local area circumstances, a feat beyond synthetic reconstruction given 1991 Census data limitations. Synthetic reconstruction is also more complex and time-consuming to program, particularly if ward level constraints are to be included. In conclusion, therefore, to generate a single set of synthetic microdata, combinatorial optimisation is greatly superior to synthetic reconstruction.

The final result of this paper has been an integrated model, called POP91, for the reconstruction of the population microdata. The structure of the model is shown in Figure 13. It contains a set of programs design to extract data, estimate the SAS 10%-based table counts, convert the data into table vector format, and use combinatorial optimisation approach to select the best households combination for each EDs. Another two tables (S43: household composition by ethnic group of head and S42b: household composition by number of car) have been added to the model, leading to a total of 16 possible constraining tables, which cover a wide range of individual and household variables in the census. The outputs are two files: one contains the selected household combinations and the other accompanying test results at both ED and ward levels, which allow users to assess the reliability of the dataset. POP91 has been used to generate a microdata for each ED (1379 in total) in Leeds metropolitan district. Adopting 14 constraining tables, the overall computer time is 33 hours on an 800 MHz PC (approximately 86 seconds per ED). Work is currently under way to extend this coverage across the UK, and to make the resulting population microdata available to users via a simple windows-based interface. Further details and progress reports may be found on the project website, http://pcwww.liv.ac.uk/~william/microdata .

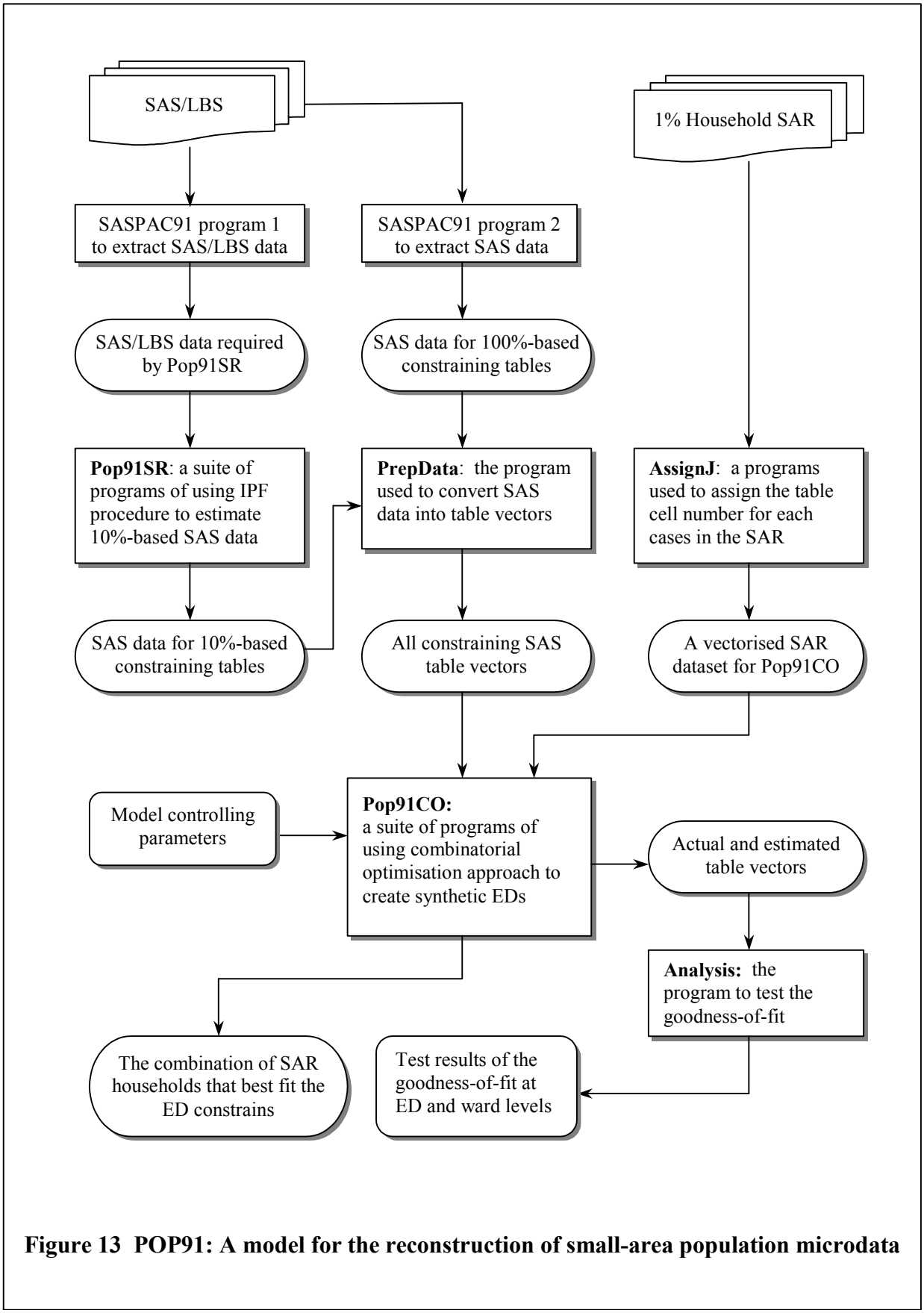


Figure 13 POP91: A model for the reconstruction of small-area population microdata

References

Birkin M and Clark M (1988) SYNTHESIS – a synthetic spatial information system for urban and regional analysis: methods and examples, *Environment and Planning A*, 20, 1645-1671.

Birkin M and Clark,G P (1995) Using microsimulation methods to synthesize census data, in Openshaw S (ed) *Census users' handbook*, GeoInformation International.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice.* Cambridge: MIT Press.

Clark,G P (1996) Microsimulation: an introduction, G Clarke (ed) *Microsimulation for urban and regional policy analysis*, Pion, London.

Dale, A. & Marsh, C., Eds. (1993) *The 1991 Census User's Guide.* London: HMSO.

Dale, A (1998) The value of the SARs in spatial and area-level research, *Environment and Planning A*, 30, 767-774.

Duley C J (1989) *A model for updating census-based household and population information for inter-censal years*. Unpublished Ph.D. Thesis, School of Geography, University of Leeds.

Fienberg S E (1970) An iterative procedure for estimation in contingency tables, *Annals of Mathematical Statistics*, 41, 349-366.

Huang Z and Williamson P (2001) A Modified Sampling Procedure for Small Area Population Simulation, *Working Paper 2001/1, Department of Geography, University of Liverpool*.

King and Bolsdon (1998) Using the SARs to add policy value to household projections, *Environment and Planning A*, 30, 867-880.

Knudsen, D.C. and Fotheringham, A.S. (1986) Matrix comparison, goodness-of-fit, and spatial interaction modeling. *International Regional Science Review,* 10, 2, 127-147.

Loh, W.-Y. and Shih, Y.-S. (1997), Split selection methods for classification trees, *Statistica Sinica*, 7, 815-840.

Norman P (1999) Putting iterative proportional fitting on the researcher's disk. Available from author at School of Geography, University of Leeds, Leeds LS2 9JT

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

Voas D and Williamson P (2000b) The scale of dissimilarity: concepts, measurement, and an application to socio-economic variation across England and Wales. *Transactions of the Institute of British Geography*, 25, 465-481.

Voas D and Williamson P (2001a) Evaluating goodness-of-fit measures for synthetic microdata, *Geographical and Environmental Modelling*, 5(2), 177-200

Voas D and Williamson P (2001b) The diversity of diversity: a critique of geodemographic classification, *Area*, 33(1), 63-76.

Williamson P (1992) *Community health care policies for the elderly: a microsimulation approach*, Unpublished Ph.D. Thesis, School of Geography, University of Leeds.

Williamson P (1993) MetaC91: a database about published 1991 Census table contents, Windows 3.1 version, Working Paper 93/18, School of Geography, University of Leeds, Leeds LS2 9JT

Williamson P (1996) Community care policies for the elderly, 1981 and 1991; a microsimulation approach, in G Clarke (ed) *Microsimulation for urban and regional policy analysis*, Pion, London, 64-87.

Williamson P (2002) Synthetic microdata. Ch 17 In *The Census Data System*, Rees P, Martin D and Williamson P (eds.). Wiley: Chichester, 231-241.

Williamson P, Rees P and Birkin M (1995) Indexing the census: a by-product of the simulation of whole populations by means of SAS and SAR data, *Environment and Planning A*, 27, 413-424

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, 785-816

Wong D W S (1992) The reliability of using the iterative proportional fitting procedure, *Professional Geographer*, 44, 340-348.