Exploring Microsimulation Methodologies for the Estimation of Household Attributes

Paper presented at the 4th International conference on GeoComputation, Mary Washington College, Virginia, USA, 25-28 July 1999

D Ballas

School of Geography, University of Leeds, Leeds LS2 9JT, UK; e-mail: D.Ballas@geography.leeds.ac.uk

G Clarke

School of Geography, University of Leeds, Leeds LS2 9JT, UK; e-mail: graham@geography.leeds.ac.uk

I Turton

School of Geography, University of Leeds, Leeds LS2 9JT, UK; e-mail: ian@geography.leeds.ac.uk

Abstract

Microsimulation is a rapidly expanding area of spatial modelling, which seems to offer great potential for applied policy analysis. However, currently there is considerable debate on the most appropriate methodology for estimating micro-data. Household or individual attribute data can be represented both as lists and/or as tabulations. It has long been argued (Birkin and Clarke, 1995; Clarke, 1996; Williamson *et al*, 1998) that the representation of information on households and individuals in the form of lists offers greater efficiency of storage and spatial flexibility as well as an ability to update and forecast. This paper reviews the possibilities and methodologies of building list-based population micro-data for small areas. First, it evaluates the methods, which have been developed and employed so far for the estimation of population micro-data, outlining the advantages and drawbacks of each one of them. Then the paper investigates the comparison of methods for generating conditional probabilities by statistical matching techniques or by using probabilities directly from household data sets such as the Samples of Anonymised Records (SARs) and the Small Area Statistics (SAS) from the UK Census of population. In addition, it explores the combination of these methods in a microsimulation framework and presents micro-data outputs from a local labour market microsimulation model for Leeds. Finally, it highlights the difficulties of calibrating this kind of model and of validating the model outputs, given the absence of suitable observed statistics.

1. Introduction

Microsimulation modelling techniques have been widely used in the social sciences during the last four decades. In particular, microsimulation modelling methodologies have been introduced by Orcutt (1957) and subsequently applied widely in Economics (see for instance Mertz, 1991; Bekkering, 1995), in Sociology (see for example Orcutt *et al*, 1986) and more recently in Geography (Clarke and Holm, 1987; Birkin and Clarke; 1988; Williamson, 1992; Clarke, 1996). Microsimulation methodologies aim at building large-scale data sets on the attributes of individuals or households and on the attributes of individual firms or organisations and at analysing policy impacts on these micro-units through the simulation of economic, demographic and social processes (Orcutt *et al*, 1986; Birkin and Clarke, 1995; Clarke, 1996).

In this paper we focus on using microsimulation techniques to build population microdata sets for small geographical areas. In particular, we build on previous work on generating population microdata and we explore different approaches to microsimulation modelling. First, in section 2 we describe the need for spatially disaggregated microdata and we outline the advantages of adopting a microdata approach. Further, in section 3 we introduce and describe different approaches to microsimulation modelling and we enumerate their principle advantages and drawbacks. In section 4 we argue the case for an object-oriented approach to microsimulation modelling and we apply different approaches to estimate labour market microdata for Leeds in an object-oriented framework. Moreover, we present model outputs and evaluate the different methods in the light of these outputs. Finally, in section 5 we give some concluding remarks and we outline future research possibilities.

2. The need for spatially disaggregated microdata

Aggregate spatial modelling has taught us much about patterns and flows within cities and regions. When combined with interesting performance indicators then we can say much about the quality of life for residents in different localities. However, we still know relatively little about the *interdependencies* between household structure or type and their lifestyles, including the events they routinely participate in and hence their ability to raise and spend various types of income and wealth. That is, we do not have a 'live' database of household types linked to earning capabilities (both earned and/or transfer payments) which can be used both to explore spatial variations in lifestyles and behaviour (the actions of individuals) and to monitor the effects of changes in taxation, family credit, pensions, social security payments etc (the actions of local and national governments). Although many countries have such information collected through national household-based censuses, very little of this is available at the household level due to confidentiality problems. We believe this presents a major challenge in contemporary quantitative geography. If we do not have a micro data base on individuals and households then there is a necessity to simulate one. Microsimulation offers various methodologies to create such micro datasets. Comprehensive reviews of microsimulation in the social sciences have been provided by Clarke and Holm (1987) and Clarke (ed, 1996). Applications have examined the

cost and distributional effects of transfer income policies and evaluated the impacts of national income and job distribution policies (Kain et al 1985, Caldwell and Kaister 1996, Holm et al 1996, Bateson et al 1980). The obvious remark to make on a large number of these models is that they are aspatial. However, it is encouraging to see that many non-geographers in the microsimulation field recognise the importance of spatial disaggregation:

> Were the locational detail in a policy model to penetrate to substate levels like county, city or even school district, such a model would find itself potentially relevant to literally hundreds and thousands of policy decisions.
>
> Caldwell (1986, p.62)

The models traditionally favoured in geography and regional science have been more aggregate meso or macro models. However, it is also increasingly recognised by many quantitative geographers that there is an urgent need to shift the focus of study to the microscale.

> To move to better dynamic representations of urban processes suggests that individuals rather than groups or aggregates must form the elemental basis of these simulations
>
> (Batty 1996, 261)

We believe that the construction of individual household data sets has a number of potential advantages over traditional approaches to urban and regional modelling (see also Clarke and Holm 1987 and Birkin and Clarke 1995). The first is computationally efficient storage. Microsimulation involves the specification of a set of attributes for a sample (which may in fact be 100%) of households and individuals that represent the population of interest. These samples, stored as lists, have a storage requirement that is simply the number of attributes multiplied by the sample size. When there is a significant number of attributes then this storage requirement will be considerably smaller than the corresponding occupancy matrix of a disaggregated meso scale model.

The second advantage relates to data linkage. There is no reason why we need to be confined to recreating the attributes from a single parent database. Provided there is a link through at least one attribute then other data sets can be included into the simulation exercise. This allows our models to be driven by new variables such as household income and expenditure. A third benefit of this approach comes from the

ability to update and forecast. The list processing approach offers a procedure ideally suited to updating. This involves examining each member of the simulated population in turn to ascertain if they are eligible for a series of transitions such a giving birth, migrating, dying etc. If an individual is eligible then the appropriate conditional probability for the transition event-taking place is used in Monte Carlo sampling procedures.

The major advantages however concern the ability to address a series of important policy questions. In this sense the proposal has a major applied policy focus. We would argue that the nature of local planning has changed in many European countries away from traditional concerns with land-use and transportation. The emphasis has shifted to a focus on the encouragement of enterprise and new business formation; improvement of job prospects through training and education; tackling dereliction and attracting investment (see Thornley (1991), Davoudi and Healey (1995), Hill (1994) amongst others). In addition, given significant changes in social policy during the Thatcher years in the UK (see Johnston 1990, Glennester 1992), it is important to recognise that so many of our day to day activities now need to be thought of as markets. Such markets have long been recognised in retailing, housing and jobs but even health and education now operate in market conditions. This places even greater emphasis on incomes and lifestyles, which are increasingly recognised to be so crucial in determining access to service provided through market mechanisms.

What then are the implications for urban and regional planning? Edwards (1995) provides a timely reminder of the importance of separating urban policy (or area-based) issues and social policy issues (which have an impact on less well off residents wherever they live). Models will have a role to play in both policy arenas. In the former there is still great interest in monitoring the impact of various types of property developments (new businesses, shops, offices etc.), and new infrastructure developments such as transport links. However, the impact assessments need to take place at much finer geographical scales. Traditional economic approaches typically produce multiplier estimations for entire cities or regions. Yet, the impacts of such developments are normally much more concentrated. What is required is a framework for modelling impacts at the household level. Similarly, if we can model the impacts and the multiplier effects of such developments we may be able to evaluate the monetary costs versus benefits of investment. The press is often littered with reports

outlining Government spending on urban policies. This may be one approach to try and evaluate the returns on this investment in terms of job creation and local multiplier effects. However, monitoring the impacts of *social* policy changes may offer a new role for urban and regional models. So far geographers and regional scientists have not been very good at addressing these issues but they may be far more important than any specific urban inner city policy:

> … by far the greatest part of the social and economic needs of inner city residents will not be met by urban-specific policies but by mainline housing, health, income support and education provision
>
> (Edwards 1995, p.701)

This is also one of Openshaw's (1995a) major agendas for his new era of computational human geography: 'Governments need to predict the outcomes of their actions and produce forecasts at the local level'. This in one sense brings up back to the classic aspatial applications of microsimulation, many of which have been designed specifically to monitor tax, income and housing benefit changes. Wegener (1986) offers a useful checklist of national policy issues which could be addressed through a more micro-scale (household) focus: taxes, regulations, credits, subsidies, government consumption, unemployment benefits, housing allowances etc. By adding the spatial dimensions through our household level modelling framework we may be able to shed new light on the local impacts of major national policy changes.

3. Microsimulation approaches to microdata generation

As aforementioned in the introduction, microsimulation is a methodology aimed at building large-scale data sets on the attributes of individuals or households and on the attributes of individual firms or organisations and at analysing policy impacts on these micro-units (Orcutt *et al*, 1986; Birkin and Clarke, 1995; Clarke, 1996). Further, microsimulation is a means of modelling real life events by simulating the social and economic characteristics and behaviour of the individual units that make up the system where the events occur. For example, microsimulation can be used to simulate the operation and effect of tax systems by applying tax rules to information about individuals, families or firms which has previously been collected

and stored on computer files[1]. Microsimulation is particularly suitable for systems where the decision-making occurs at the individual unit level and where the interactions within the system are complex. In such systems, the outcomes produced by altering the system can vary considerably and widely for different groups and are often difficult to predict. Because microsimulation models are concerned with the behaviour of micro-units (such as households or firms) they are especially well suited to estimate and analyse the distributional impacts of policy changes (Mertz, 1991). In addition, microsimulation modelling frameworks provide the possibility of defining the goals of economic and social policy, the instruments employed and also the structural changes of those affected by socio-economic policy measures (Krupp, 1986).

Microsimulation modelling requires the construction of a microdata set. In particular, microsimulation modelling aims at building a population of microunits along with their associated characteristics. This procedure usually involves the use of contingency tables or conditional probability analysis to estimate chain conditional probabilities. In particular, conditional probabilities are calculated from available known data and then they are used to reconstruct detailed micro-level populations. This process can be clarified with the use of an illustrative example. Lets assume that we wish to investigate the relationships between sex (*S*), age (*A*), educational qualifications (*Q*), economic position (*EP*) and socio-economic group (*SEG*) for a given population group X in location *i*. From the Small Area Statistics (SAS) tables of the 1991 Census of the UK population we can obtain for the population in a specified area (e.g. at the ward level) separate tabulations of:

- sex by age by economic position (SAS table 08)
- level of qualifications by sex (SAS table 84)
- socio-economic group by economic position (SAS table 92)

From these tabulations we could calculate the respective conditional probabilities, and then our problem would be to estimate the probability:

- $p(x_i, S, A, Q, EP, SEG)$

given a set of constraints or known probabilities:

- $p(x_i, S, A, EP)$

[1] http://natsem.canberra.edu.au/html/microsimulation.html

- $p(x_i, Q, S)$
- $p(x_i, SEG, EP)$

There are a number of ways to solve this problem such as linear programming models, discrete choice models, balancing factor methods in spatial interaction models (Wilson, 1970) and Iterative Proportional Fitting techniques (Birkin, 1987; Clarke, 1996; Wong, 1992; Fienberg, 1970).

The second stage of the microsimulation procedure is to create a sample of individuals based on this set of probabilities. The creation of such a data set can be achieved by Monte Carlo simulation. Clarke (1996) illustrates how this procedure can be employed for the creation of a micro-level population with the following characteristics: age, sex, marital status and household tenure (see figure 1). Supposing that the age, sex, and marital status of the head of household, is available from the Census, it is possible, using a statistical matching procedure, to estimate probabilities of household tenure given head of household's age, sex, and marital status (box 2 in figure 1). The first synthetic household in this example has the following characteristics: male head of household, aged 27, married. As it can be seen in box 2 the estimated probability that a household of this type would be owner-occupied is 70%. The next step in the procedure is to generate a random number to see if the synthetic household gets allocated to the owner occupied category. The random number in this example is 0.542 and falls within the 0.001 - 0.700 range needed to qualify as owner-occupied. The same procedure is then carried out sequentially for the tenure allocation to all the synthetic households (Clarke, 1996). It should be noted that the difficult task in microsimulation is to specify which variables are dependent upon others and to determine the ordering of probabilities (Ibid.).

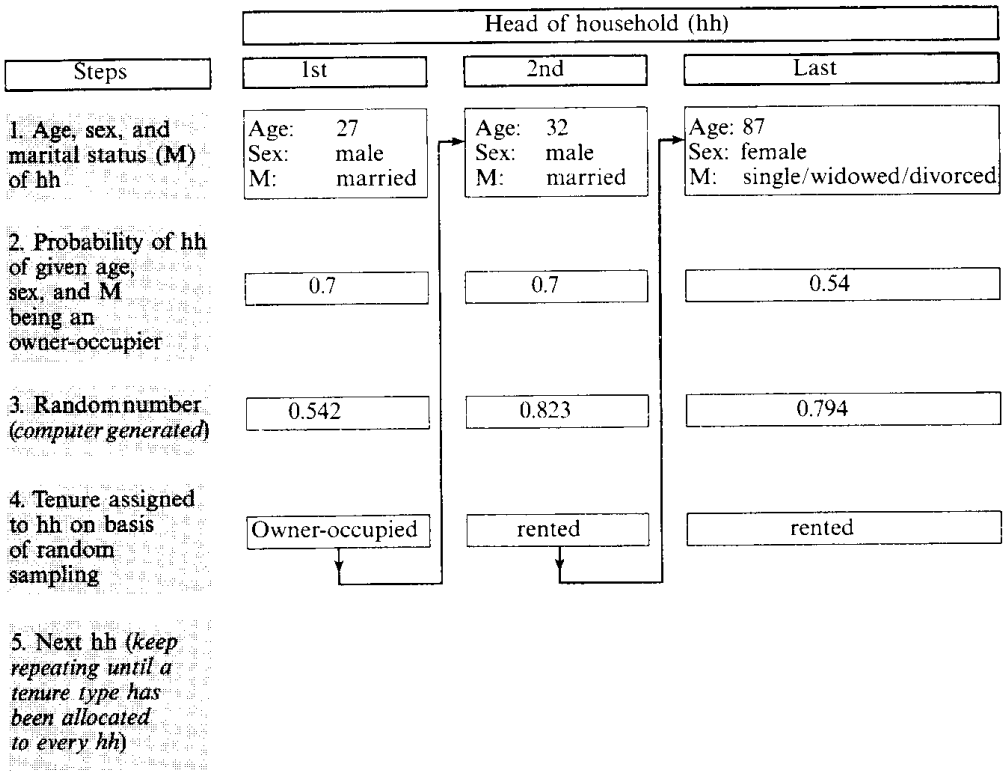


Figure 1: An example of the microsimulation process: tenure allocation procedure (after Clarke, 1996:3)

A different approach to microdata generating is through the reweighting of a parent sample of microdata, which is available at a different (from the desired) spatial scale. This procedure involves the use of a sample of microunits which is adjusted in such a way as to be representative of a parent population as represented in mesoscale or microscale tables (Clarke, 1996; Williamson *et al*, 1998). One of the merits of this approach is that it exploits existing microdata sets, which are available at regional or national scales such as the Samples of Anomymised Records in the UK (see Marsh and Teague, 1992; Middleton, 1995) and the Public Use Microdata Samples (PUMS) in the USA.

A reweighting approach to micro-population data generation has been adopted by Williamson *et al* (1998) who explored different solutions to finding the combination of household SAR records which best fit known small area constraints. In order to deal with the above problem they used various techniques of combinatorial optimisation such as hill climbing algorithms, simulated annealing approaches and

genetic algorithms. After thoroughly testing these approaches, they concluded that a modified simulated annealing algorithm stands out as the best solution.

Table 1 provides a useful outline of the merits and drawbacks of microsimulation modelling. Detailed discussions of these of these advantages and disadvantages can be found in the literature (see for example Clarke and Holm, 1987; Birkin and Clarke, 1995; Clarke, 1996; Williamson *et al*, 1998).

Advantages	*Drawbacks*
Data linkage	***Difficulties in calibrating the model and validating the model outputs***
Spatial flexibility	***Large requirements of computational power***
Efficiency of storage	
Ability to update and forecast	

Table 1: Advantages and drawbacks of microsimulation

In the remainder of this paper we discuss different microsimulation approaches to generating population microdata. First, we argue the case for an object-oriented approach to microsimulation modelling. We then outline different methodologies of generating conditional probabilities and we present 5 different object-oriented microsimulation models that apply these methodologies.

4. An object-oriented approach to microsimulation modelling

4.1 The rationale for building object oriented microsimulation models

As seen above, microsimulation frameworks have the advantages of a list-based approach to microdata representation. In such frameworks variables are treated as lists rather than matrices. For instance, in the context of a microsimulation approach a household has a list of attributes which are all stored as lists rather than as occupancy matrices (Clarke, 1996). From a computer programming perspective, the household can also be seen as an object with its associated object variables (attributes). Therefore, it can be argued that an *object-oriented* language such as *JAVA* or *C++* is most suitable for microsimulation modelling. It should be noted that *JAVA* has the added advantage of platform independence, which allows for the execution of microsimulation programs in the World Wide Web.

In an *object-oriented* system, a *class* is a collection of *data* and *methods* that operate on that data. Further, the data and methods describe the state and behaviour of

an *object* (Flanagan, 1997; Lemay and Perkins, 1997). Classes can also be seen as templates for multiple objects with similar features. They embody all the features of a particular set of objects. Hence, we can have a *household* class that describes the features of all households (e.g. age, sex and marital status of head of household, tenure, etc.) The *household* class serves as an abstract model for the concept of a household. Once a *household class* is defined then lots of different *instances* of that class can be created and each different *household instance* can have different features while still being immediately recognisable as a household.

In the context of this paper we designed a *household class* with the variables listed in table 12.

Micro-household attributes
Broad age group of head of household (HH)
Sex of HH
Marital status of HH
Tenure
Economic activity of HH
Employment status of HH
Ethnic group of HH
Occupation of HH
Industry of HH
Educational qualifications of HH
Hours worked of HH
Former industry of unemployed of HH

Table 12 – The variables of class Household

The Household class contains methods that operate on the variables defined. In particular, these methods require the input of conditional probability rates and invoke the Monte Carlo procedure to assign household attributes on the basis of random sampling in a similar manner to that depicted in figure 1. The method *random* of the JAVA class *Math* is employed to generate a random number greater than or equal to 0.0 and less than 1.0. The implementation of this method uses the *java.util.Random* class (Grand and Krunsden, 1997). This class is a *pseudo-random number generator* that generates pseudo-random numbers by starting with a seed value and then using an algorithm to generate a sequence of numbers that appear to be random. The *Random*

class uses a 48-bit seed and a *linear congruential algorithm* to change the seed (for more details see Grand and Krunsden, 1997).

Having created a Household class the next step is to *instantiate* the class, that is to create *Household objects* or, in other words, create *instances* of the class *Household*. In order to do so we needed a baseline population and probabilities for every household attribute, conditional upon the attributes of the baseline population. As seen in the previous section there are various methodologies for the estimation of conditional probabilities. In the next subsections we apply these methodologies and we present five different microsimulation models which *instantiate* the class Household. The first model uses SAS generated probabilities and performs the *Monte Carlo sampling* procedure to assign attributes to a baseline population. The second model performs Monte Carlo sampling from conditional probability distributions generated from the SARs, while the third model uses probabilities generated from the use of IPF on data from the SAS and the SARs. Finally, the fifth model adopts a reweighting approach to microsimulation modelling.

It can be argued that the first three models adopt a *bottom-up* micro-modelling *approach*, whereas the fourth and the fifth model follow a *top-down* micro-modelling approach.

4.2 SimLeeds1: using conditional probabilities from the Small Area Statistics

4.2.1 Using the Small Area Statistics (SAS) tables to calculate probabilities

One way of deriving probabilities that can be used in microsimulation modelling is to extract these probabilities from each census table, which is related to the modelling task. For example, if we wished to combine the Small Area Statistics (SAS) tables 8 (Economic activity) and 39 (age, sex and marital status of head of household) we could use the latter to construct an initial baseline population and the former to calculate the probabilities of different population groups having different economic activity attributes. Then, we could use these probabilities to assign economic activity characteristics to each head of household of table 39.

An example is useful at this point to clarify the process. Table 2 depicts the numbers of heads of households by age, sex, and marital status in the first Enumeration District (ED) of Beeston in Leeds, while tables 3 and 4 present the

counts of male and female residents, respectively, in this ED by economic activity category.

	Total persons	Total males	Males swd	Males married	Total females	Females swd	females married
All ages	235	162	97	65	73	67	6
16-29	55	40	33	7	15	13	2
30-44	64	50	34	16	14	14	0
45-59	55	34	15	19	21	18	3
60-64	19	16	7	9	3	3	0
65-74	23	14	5	9	9	9	0
75-84	17	7	3	4	10	9	1
85+	2	1	0	1	1	1	0

Table 2: Heads of households by age, sex and marital status, ED1, Beeston, Leeds

Source: The 1991 Census, Crown Copyright. ESRC purchase.

From tables 3 and 4 we can derive the probabilities of each resident to belong to each economic activity category. Thus, tables 3 and 4 depict these probabilities which can be used for microsimulation modelling.

Economically Active males	Total 16+	16-29	30-44	45-59	60-64	65+
total males	209	74	59	38	15	23
total economically active males	161	70	54	30	6	1
Full time employees	104	50	33	17	4	0
Part time employees	5	1	0	2	1	1
Self-employed with employees	1	0	0	1	0	0
Self-employed without employees	9	2	5	2	0	0
On a Government scheme	6	2	3	1	0	0
Unemployed	36	15	13	7	1	0

Table 3: Economically active males, ED 1, Beeston, Leeds

Source: The 1991 Census, Crown Copyright. ESRC purchase.

Economically Active females	Total 16+	16-29	30-44	45-59	60-64	65+
total females	167	53	32	37	13	32
total economically active females	79	39	17	19	2	2
Full time employees	43	28	10	5	0	0
Part time employees	23	4	6	9	2	2
Self-employed with employees	1	0	0	1	0	0
Self-employed without employees	0	0	0	0	0	0
On a Government scheme	1	0	0	1	0	0
Unemployed	11	7	1	3	0	0

Table 4: Economically active females, ED1, Beeston, Leeds

Source: The 1991 Census, Crown Copyright. ESRC purchase.

Probabilities (males)	Total 16+	16-29	30-44	45-59	60-64	65+
Full time employees	0.646	0.714	0.611	0.567	0.667	0.000
Part time employees	0.031	0.014	0.000	0.067	0.167	1.000
Self-employed with employees	0.006	0.000	0.000	0.033	0.000	0.000
Self-employed without employees	0.056	0.029	0.093	0.067	0.000	0.000
On a Government scheme	0.037	0.029	0.056	0.033	0.000	0.000
Unemployed	0.224	0.214	0.241	0.233	0.167	0.000

Table 5: Economic activity probability rates, ED 1, Beeston, Leeds

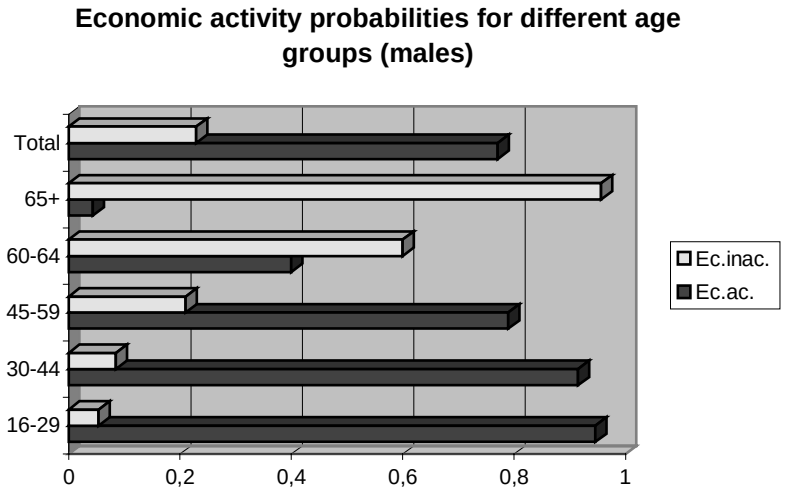
Source: The 1991 Census, Crown Copyright. ESRC purchase.

Probabilities (females)	Total 16+	16-29	30-44	45-59	60-64	65+
Full time employees	0.544	0.718	0.588	0.263	0.000	0.000
Part time employees	0.291	0.103	0.353	0.474	1.000	1.000
Self-employed with employees	0.013	0.000	0.000	0.053	0.000	0.000
Self-employed without employees	0.000	0.000	0.000	0.000	0.000	0.000
On a Government scheme	0.013	0.000	0.000	0.053	0.000	0.000
Unemployed	0.139	0.179	0.059	0.158	0.000	0.000

Table 6: Economic activity probability rates, ED1, Beeston, Leeds

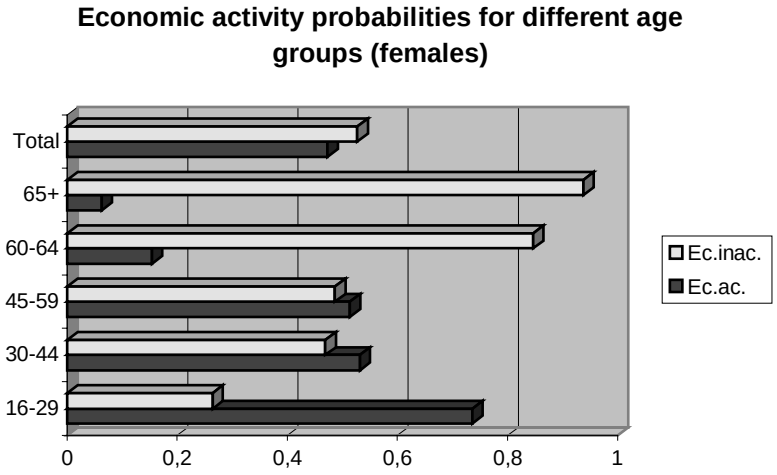
Source: The 1991 Census, Crown Copyright. ESRC purchase.

In addition, from table 8 of the SAS we can calculate for different age groups the probabilities of being economically active or inactive (see graphs 1 and 2).



Graph 1

Source: The 1991 Census, Crown Copyright. ESRC purchase.



Graph 2

Source: The 1991 Census, Crown Copyright. ESRC purchase.

The data from the above tables and graphs can be used to construct a micropopulation of heads of households with the following attributes:

- age
- sex
- marital status
- economic activity or inactivity
- economic activity category

More specifically, the data from table 3 provide an initial population of heads of households at the ED level. In addition, following the procedure, which was shown in Figure 1, and using the probabilities derived from table 8 of the SAS we can assign economic activity status to each head of household. The same methodology may be followed in order to construct a micropopulation with the characteristics of table 2.

However, it should be noted that the above approach has an important drawback. In particular, the probabilities derived from a single SAS table are not conditional upon all the desired characteristics of the household or the individual. More specifically, in the above example, the economic activity rates calculated from the SAS table 8 are conditional upon the age, sex and the geographical location (at the ED level) of the head of household, but they are not conditional upon marital status. This problem can be overcome with the use of the Samples of Anonymised Records for the probability calculation.

4.2.2 Monte Carlo sampling from SAS generated conditional probabilities

This section presents *SimLeeds1*, which was built in JAVA to estimate the micro-population attributes of heads of households listed in table 12. The model uses conditional probabilities extracted from the SAS census tables and estimates the head of household attributes on the basis of random sampling. There are two classes in this model: *Household* and *Llmsim*. Also, there is class *SimLeeds* that runs as an application to apply classes *Household* and *Llmsim*.

The variables of class Household represent the attributes of table 12. In addition, class *Household* has several *methods*, which require a set of conditional probabilities and invoke Monte Carlo sampling to estimate the missing attributes. For example the *Household* method *getEcac(double p)* requires the conditional probability *p* of a given head of household to be economically active and it assigns economic activity status to this head of household on the basis of random sampling. Likewise, methods *getEthnicity(double p[ ])* and *getTenure(double p[ ])* require a set of conditional probabilities for the ethnicity category and tenure of a given head of household. Table 13 contains all the methods of class *Household*.

Class *Llmsim* has two variables. The first variable *popcon* contains the *population constraints* or *baseline population*. The Small Area Statistics table 39 provides this baseline population. SAS table 39 provides the counts of heads of

households by age (7 age groups), sex, and marital status. Therefore, there are 7 x 2 x 2 = 28 different initial household categories.

Table 13 - Methods of class Household
`getEcAc(double p)`
`GetEcAcCat(double p[])`
`GetEthnicity(double[] p)`
`GetTenure(double[] p)`
`GetOccupation(double[] p)`
`GetIndustry(double[] p)`
`GetEducation1(double p)`
`GetEducation2(double p[])`
`GetHours(double[] p)`
`GetFormerIndustry(double[] p)`

The second variable of *Llmsim* is designed to store all the conditional probabilities, which are required to assign more attributes to the baseline population. This variable is defined as a *4-dimensional array* that holds all the conditional probability values, which are required for the microsimulation modelling. In particular, the elements of this array are associated with the following files:

- Probability files for economic activity status - these files contain probabilities of economic status (i.e. active or inactive) conditional upon age group and sex. In particular, it contains the probabilities of heads of household of each age and sex category to be economically active for each males and females respectively The probabilities for this files are derived from SAS table 8. [2]
- Probability files for employment status -these files contain probabilities of employment status (i.e. employee, self-employed, on a government scheme, unemployed) conditional upon age group and sex for males and females respectively. The probabilities for these files are derived from SAS table 8.
- Probability file for ethnicity - This file contains probabilities of ethnicity conditional upon sex derived from SAS table 6.
- Probability file for occupation - this file contains occupation probabilities conditional upon sex and employment status, which are derived from SAS table 81.
- Probability file for industry - this file contains industry probabilities conditional upon employment status and sex and they that are derived from SAS table 79.
- Probability file for educational qualifications - this file contains educational qualifications' probabilities conditional upon sex (derived from SAS table 84).

[2] An index with all the SAS tables can be found in Openshaw (ed.), 1995.

- Probability file for industry - this file contains the *hours-worked* probabilities for each age group conditional upon sex (derived from SAS table 79).
- Probability file for former industry of unemployed - this file contains former industry probabilities of unemployed males and females (derived from SAS table 94).
- Probability file for tenure - this file contains tenure probabilities (derived from SAS table 87).

Class *Llmsim* has several input methods, which read the above probability files and store the probability values in the array variable of *Llmsim*. In addition, *Llmsim* has the method *simulateHouseholds* to simulate households using the above conditional probabilities, which are provided as input to the program, and the *Monte Carlo* methods of class *Household*.

As mentioned above, the class *SimLeeds* operates as a program and applies the classes *Household* and *Llmsim*. The latter uses the baseline population data to generate an array with Household objects. It then uses the method *simulateHouseholds* to simulate all the characteristics of table 12 for each household object. More specifically, the program follows the steps below:

- **Step 1** - Read the baseline population and the probability files (generated from the SAS)
- **Step 2** - Create an array of household objects based on the baseline population data
- **Step 3** - Assign each household the characteristics of table 12 on the basis of random (Monte Carlo) sampling, using the conditional probability distributions given in the probability files.
- **Step 4** - Output the list of microsimulated households

As can be seen, *SimLeeds* generates a list of the estimated households. However, it should be noted that since there are no available micro-data sets at this level of geographical scale (ED level), the only way to validate the analysis is to derive cross-tabulations of variables from the estimated list of micro-population which match published census cross-tabulations. In addition, since different tables have been used (i.e. tables referring to individuals and tables referring to heads of households) it is not meaningful to compare the cross-tabulations of the estimated variables with the

published census cross-tabulations in the form of counts. Instead, a comparison of rates is much more suitable. In addition, since the model is based on a random sampling procedure there is a considerable degree of variation in the outputs. Therefore, it is useful to run the model several times and take the average of all the outputs as a more reliable estimate. For the purposes of this paper we ran the *SimLeeds1* model five times and derived the five-run averages of all the estimates. Table 14 depicts census cross-tabulations of economic activity and age for males in the form of rates, while table 15 depicts cross-tabulations of the respective estimated rates (five-run average). In addition, table 16 presents the total absolute difference between the observed and estimated rates (Total Absolute Error).

Age	Active	Inactive
16-29	0.946	0.054
30-44	0.915	0.085
45-59	0.789	0.211
60-64	0.400	0.600
65+	0.043	0.957
Total	0.570	0.430

Table 14 - Economic status observed probability rates of males, ED 1, Beeston, Leeds

Source: The 1991 Census, Crown Copyright. ESRC purchase.

Age	Active	Inactive
16-29	0.925	0.075
30-44	0.856	0.144
45-59	0.782	0.218
60-64	0.500	0.500
65+	0.045	0.955
Total	0.622	0.378

Table 15 - Economic status estimated probability rates of males, ED 1, Beeston, Leeds

Age	Active	Inactive
16-29	0.021	0.021
30-44	0.059	0.059
45-59	0.007	0.007
60-64	0.100	0.100
65+	0.002	0.002

Table 16 - Economic status TAE – absolute difference between estimated probability rates of males, ED 1, Beeston, Leeds

Likewise, table 17 presents a census cross-tabulation of employment status rates and age for females, whereas table 18 gives the estimated rates derived from *SimLeeds1* (five-run average) and table 19 gives the total absolute difference between the observed and estimated rates.

Age	Employees	Self employed	On a Government scheme	Unemployed
16-29	0.821	0.000	0.000	0.179
30-44	0.941	0.000	0.000	0.059
45-59	0.737	0.053	0.053	0.158
60-64	1.000	0.000	0.000	0.000
65+	1.000	0.000	0.000	0.000

Table 17 - Employment status observed probability rates of females, ED 1, Beeston, Leeds

Source: The 1991 Census, Crown Copyright. ESRC purchase.

Age	Employees	Self employed	On a Government scheme	Unemployed
16-29	0.818	0.000	0.000	0.182
30-44	0.952	0.000	0.000	0.048
45-59	0.857	0.036	0.089	0.018
60-64	1.000	0.000	0.000	0.000
65+	1.000	0.000	0.000	0.000

Table 18 - Employment status estimated probability rates of females, ED 1, Beeston, Leeds, SimLeeds1 (five-run average)

Age	Employee	Self-employed	On a Government scheme	Unemployed
16-29	0.002	0.000	0.000	0.002
30-44	0.011	0.000	0.000	0.011
45-59	0.120	0.017	0.037	0.140
60-64	0.000	0.000	0.000	0.000
65+	0.000	0.000	0.000	0.000

Table 19 - Employment status TAE – absolute difference between estimated probability rates of females, ED 1, Beeston, Leeds. SimLeeds1 (five-run average)

TAE statistics	First run	Second run	Third run	Fourth run	Fifth run	5-run average
Average	0.059	0.071	0.069	0.134	0.017	0.038
Max	0.100	0.225	0.213	0.180	0.029	0.100
Min	0.002	0.002	0.002	0.088	0.002	0.002
SUM	0.595	0.706	0.695	1.338	0.171	0.379

Table 20 - Economic status –TAE for all runs and for 5-run average (males).

TAE statistics	First run	Second run	Third run	Fourth run	Fifth run	5-run average
Average	0.156	0.124	0.134	0.100	0.104	0.057
Max	0.513	0.154	0.180	0.179	0.201	0.179
Min	0.010	0.088	0.088	0.010	0.003	0.003
SUM	1.556	1.243	1.338	0.999	1.045	0.566

Table 21 - Economic status –TAE for all runs and for 5-run average (females).

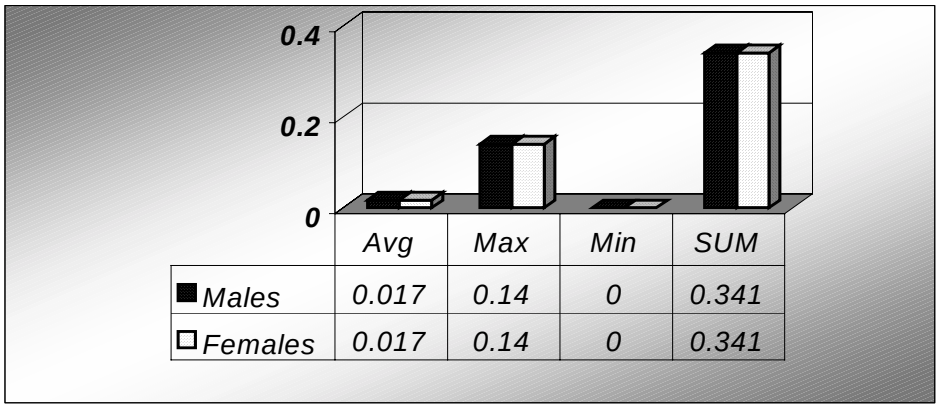
TAE statistics	First run	Second run	Third run	Fourth run	Fifth run	5-run average
Average	0.037	0.030	0.044	0.039	0.027	0.017
Max	0.167	0.160	0.167	0.180	0.144	0.140
Min	0.000	0.000	0.000	0.000	0.000	0.000
SUM	0.737	0.607	0.878	0.780	0.537	0.341

Table 22 - Employment status –TAE for all runs and for 5-run average (males).

TAE statistics	First run	Second run	Third run	Fourth run	Fifth run	5-run average
Average	0.032	0.092	0.083	0.086	0.027	0.017
Max	0.158	1.000	1.000	1.000	0.196	0.140
Min	0.000	0.000	0.000	0.000	0.000	0.000
SUM	0.641	1.833	1.669	1.716	0.543	0.341

Table 23 - Employment status –TAE for all runs and for 5-run average (females).

Further, tables 20-23 present a comparison of the TAE between the five runs and the five-run average. As can be seen, the five-run average has the lowest TAE. Graph 3 summarises the five-run average TAE statistics for males and females



Graph 3 Employment status – TAE statistics for males and females, SimLeeds1 (five-run average)

Nevertheless, as it has been pointed out in the previous sections, the above modelling approach has the important drawback of not using probabilities, which are conditional upon all the known characteristics. In the next section we present a model which deals with this problem by using conditional probabilities derived from the Samples of Anonymised Records (SARs).

4.3 SimLeeds2: Using probabilities from the Samples of Anonymised Records (SARs)

4.3.1 Using the Samples of Anonymized Records (SARs) to calculate probabilities

The 1991 Census of the UK population was the first to include the Samples of Anonymised Records (SARs), which is a major new source of census information about people in Britain (Marsh, 1993; Middleton, 1995). These are samples of individual census records which are anonymized in various ways to ensure that there is no breach of the confidentiality of the census and that no individual can be identified from the data (Middleton, 1995). The SARs is a valuable and very important source of data for the validation of model outputs and estimates. In

particular, spatially disaggregated probability distributions can be spatially aggregated at the district, regional or national level and be compared to the probability distributions derived from the SARs.

Nevertheless, the usefulness of the SARs, from a computer modelling viewpoint, is not limited to model output validation. In contrast, the SARs can be used as inputs to modelling procedures. In particular, the SARs provide an alternative source of data to be used for the calculation of probabilities. In relation to the example of the previous section, the SARs (instead of SAS tables) could be used in order to estimate the economic activity probabilities of certain categories of heads of households. Using the table 39 of the SAS for the provision of an initial baseline population we can calculate from the SARs the economic activity and status probabilities conditional upon age, sex and marital status of head of household and then incorporate these probabilities in the microsimulation procedure. Likewise, we can calculate conditional probabilities for additional household attributes in order to estimate them.

4.3.2 Monte Carlo sampling from SARs generated conditional probabilities

SimLeeds2 has the same class structure with the first model. Nevertheless, *SimLeeds2* uses conditional probabilities extracted from the SARs and estimates the head of household attributes on the basis of random sampling. As in the previous model, there is a *Household* class and a *Llmsim* class. Also, there is the class *SimLeeds*, which runs as an application to apply classes *Household* and *Llmsim*. Further, the variables and methods of classes *Household* and *Llmsim* are the same as in the previous section.

The baseline population data are also the same and they are used by classes *Llmsim* and *SimLeeds* to generate an array with Household objects. Method *simulateHouseholds* is then employed to simulate all the characteristics of table 12 for each household object. More specifically, the *SimLeeds2* follows the steps below:

- **Step 1** - Read the baseline population and the probability files (generated from the SARs)
- **Step 2** - Create an array of household objects based on the baseline population data

- **Step 3** - Assign each household the characteristics of table 12 on the basis of random (Monte Carlo) sampling, using the conditional probability distributions given in the probability files.
- **Step 4** - Output the list of microsimulated households

As it was the case with *SimLeeds1*, *SimLeeds2* generates a list of the estimated households. As far as model output validation is concerned, the same problem arises here, as there are no available micro-data sets at the ED level. Therefore, in order to validate the analysis we need to derive cross-tabulations of variables from the estimated list of micro-population, which match published census cross-tabulations. Also, as in the previous model, we have used probability data to assign characteristics to heads of households. However, the published Census cross-tabulations of these characteristics refer to individuals and therefore a comparison of rates is much more suitable rather than comparing counts.

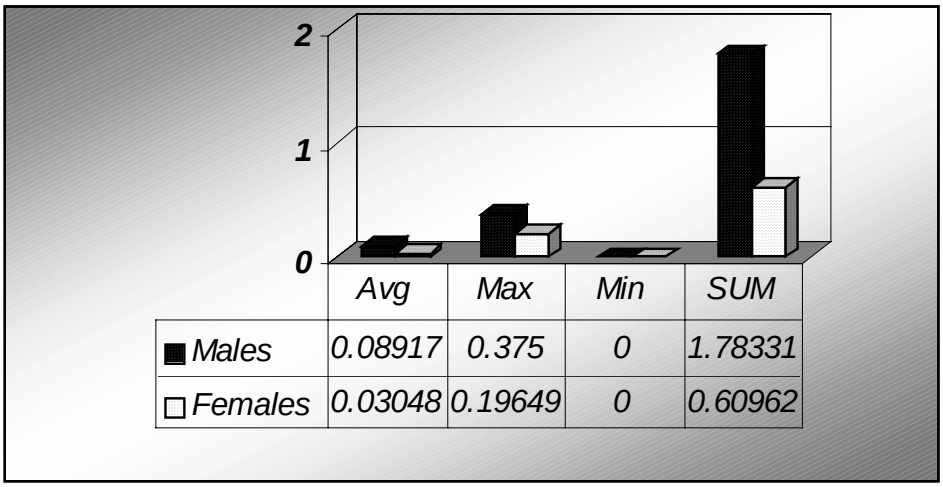
In addition, since different tables have been used (i.e. tables referring to individuals and tables referring to heads of households) it is not meaningful to compare the cross-tabulations of the estimated variables with the published census cross-tabulations in the form of counts. Instead, a comparison of rates is more appropriate. Further, as it is the case with all the microsimulation models, *SimLeeds2* is based on random sampling procedures and, hence, there is a considerable degree of variation in the outputs. Thus, *SimLeeds2* was ran five times and five-run averages have been calculated for all the estimates. Tables 24 and 25 present the TAE for these estimates whereas graph 3 depicts some TAE statistics.

Age	Employee	Self-employed	On a Government scheme	Unemployed
16-29	0.009	0.058	0.007	0.042
30-44	0.098	0.048	0.020	0.126
45-59	0.145	0.007	0.033	0.119
60-64	0.073	0.160	0.000	0.087
65+	0.375	0.375	0.000	0.000

Table 24 - Employment status TAE – absolute difference between estimated probability rates of males, ED 1, Beeston, Leeds, SimLeeds2 (five-run average)

Age	Employee	Self-employed	On a Government scheme	Unemployed
16-29	0.001	0.020	0.040	0.059
30-44	0.048	0.018	0.000	0.030
45-59	0.196	0.039	0.053	0.105
60-64	0.000	0.000	0.000	0.000
65+	0.000	0.000	0.000	0.000

Table 25 - Employment status TAE – absolute difference between estimated probability rates of females, ED 1, Beeston, Leeds, SimLeeds2 (five-run average)



Graph 4 - Employment status – TAE statistics for males and females, SimLeeds2 (five-run average)

Looking at the above graph and in comparison to graph 3, we see that *SimLeeds2* has a larger average TAE for both males and females. It should be noted though that one of the big disadvantages of using conditional probabilities derived from the SARs is that these are not conditional upon the geographical location at the ED level, as the household microdata from the SARs are released only at regional or metropolitan district level.

It can be argued that more robust estimates of conditional probabilities at the ED level can be produced with the employment of methods such as the Iterative Proportional Fitting, which is discussed below.

4.4 SimLeeds3 and SimLeeds4: Monte Carlo sampling from SAS generated conditional probabilities with the use of IPF vs Monte Carlo Sampling from SAS and SARs generated conditional probabilities with the use of IPF

4.4.1 The Iterative Proportional Fitting (IPF) technique

4.4.1.1 Definition

The Iterative Proportional Fitting (IPF) technique is used to overcome data shortfalls, where there is a lack of detail at some spatial or other level in source data (Williamson *et al*, 1996). IPF is a function which requires many repetitions to derive a probability distribution which is itself constrained by other distributions (Birkin, 1986; Fienberg, 1970; Williamson *et al*, 1996; Williamson, 1992). In particular, the

IPF technique is based on contingency table analysis and it appears in a variety of forms, from balancing factors in spatial interaction modelling through to the RAS method in economic accounting (Clarke, 1996; Birkin and Clarke, 1988). The IPF procedure can be seen in the simplest case as a method to adjust a two-dimensional matrix iteratively until the row sums and column sums equal some predefined values (Birkin, 1987; Wong, 1992). IPF can also be defined as a mathematical scaling procedure, which ensures that a two-dimensional table of data is adjusted so that its row and column totals agree with row and column totals from alternative sources (Norman, 1999) or it can be seen, from a geographical viewpoint, as a procedure for generating disaggregated spatial data from spatially aggregated data (Wong, 1992). Formally, the IPF procedure can be expressed as follows:

$$p_{ij(k+1)} = \frac{p_{ijk}}{\sum_j p_{ijk}}$$

$$p_{ij(k+2)} = \frac{p_{ij(k+1)}}{\sum_j p_{ij(k+1)}} \times Q_j$$

Where p is the matrix to be adjusted, p_{ijk} is the matrix element in row i, column j and iteration k, Qi and Qj are the predefined row sums and column sums respectively. Equations (1) and (2) are employed iteratively to estimate the new cell values until iteration m where:

$$\sum_j p_{ijm} = Qi$$

and

$$\sum_i p_{ijm} = Qj$$

Illustrative examples of the IPF procedure can be found in Wong (1992), Rees (1994) and Norman (1999). Also, Wong (1992) addresses the reliability of the method and evaluates the importance of different factors, which affect the performance of the IPF procedure. The mathematical properties of IPF and the theory underpinning it has been discussed in more detail elsewhere (see for instance Fienberg, 1970; Birkin 1987; Birkin and Clarke, 1988a). In the following paragraphs we demonstrate how

IPF can be applied on data from the 1991 UK Census to compute conditional probabilities for microsimulation modelling.

4.4.1.2 Applying IPF on data from the Small Area Statistics to calculate conditional probabilities for microsimulation modelling

One way of employing IPF to compute conditional probabilities for spatial microsimulation modelling is to apply it on data from the Small Area Statistics. In particular, with reference to the example used in section 3.2, it is possible to employ IPF to estimate the joint probability distribution of the following household attributes:

- age
- sex
- marital status
- employment status

As seen in section 4.2.1 the SAS table 39 contains data from which it is possible to derive the probability of a Head of Household being *male* or *female* and *married* or *Single-Widowed-Divorced* conditional upon *age* and *location* (at the ED level). It is possible to derive from SAS table 8 the respective conditional probabilities for *economic activity* and *employment status*. Table 7 depicts the marital status probability rates for male heads of households aged 16-29 years for the first ED of Beeston, while table 9 contains the economic activity rates for male individuals of the same age for the same location.

Marital status	SWD	Married
Males aged 16-29	0.825	0.175

Table 7: Probability rates for marital status, first ED of Beeston, Leeds

Source: The 1991 Census, Crown Copyright. ESRC purchase.

	Econ. Inactive	Employee	Self employed	On a government scheme	Unemployed
Males aged 16-29	0.054	0.689	0.027	0.027	0.202

Table 8: Probability rates for employment status, first ED of Beeston, Leeds

Source: The 1991 Census, Crown Copyright. ESRC purchase.

The IPF procedure can be applied on the data from tables 7 and 8 to estimate the joint probability distribution:

p(location,age,sex,marital-status, employment-status)

More specifically, as it can be seen in table 9, the IPF procedure can be employed to estimate the missing cell values (which, in this table, take a starting value of 1).

Household attributes	**Econ. Inactive**	**Employee**	**Self employed**	**On a government scheme**	**Unemployed**	**Row constraint**
SWD	1	1	1	1	1	0,825
Married	1	1	1	1	1	0,175
Column constraint	0,054	0,689	0,027	0,027	0,202	Σ = 1

Table 9 – Probability matrix and constraints before applying the IPF

Within table 9 there is a matrix with all its cell values set to 1. This matrix depicts the unknown distribution

p(location,age,sex,marital-status, employment-status)

which is constrained to the known probability distributions of marital status (column constraints) and employment status (row constraints). Table 10 shows the estimated values for the above probability distribution after the completion of the IPF procedure.

Household attributes	**Econ. Inactive**	**Employee**	**Self employed**	**On a government scheme**	**Unemployed**	**Row constraints**
SWD	0.044	0.568	0.022	0.022	0.167	0.825
Married	0.009	0.120	0.004	0.004	0,035	0.175
Column constraints	0.054	0.689	0.027	0.027	0.202	Σ = 1

Table 10 – Estimated probability distribution with IPF for the first ED of Beeston, Leeds

As it can be seen, the original matrix has been adjusted so that its row and column sums match the row and column constraints respectively. It should be noted though that, in this example, there is no need for more than one iteration because there are relatively few categories in the variables and there is no information about the 'interaction' between the cells of the estimated matrix. In cases where such information is available it can be input in the IPF procedure.

IPF can be applied on a combination of data from more SAS tables, in order to estimate probability distributions for a larger number of household attributes. In addition, it can be carried out separately for every spatial zone (e.g. for each ED).

4.4.1.3 Applying IPF on data from the Samples of Anonymised Records (SARs) and the Small Area Statistics (SAS) to calculate conditional probabilities for microsimulation modelling

Another way of using IPF to derive conditional probabilities is to combine data from the SAS and the SARs. As mentioned above, the SARs have a very crude geographical reference. One way for spatially disaggregating probability distributions derived from the SARs is to apply the IPF procedure on probability data from the SARs and the SAS. Table 11 gives an illustrative example of how data from the SAS and the SARs can be used to spatially disaggregate the SARs distributions at the ED level with the use of IPF.

Household attributes / zone	**Econ. Inactive**	**Employee**	**Self employed**	**On a government scheme**	**Unemployed**	**SAS row constraint**
DAFA01	0.4	0	0	0	0.6	0.014621
DAFA02	0.47368	0	0	0	0.526316	0.058484
DAFA03	0.63636	0	0	0.090909	0.272727	0.043863
DAFA04	0.61538	0	0	0.076923	0.307692	0.058484
DAFA05	0.22222	0	0.111111	0	0.666667	0.021931
DAFA06	0.34375	0	0	0.03125	0.625	0.00731
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
DAGK45	0.90909	0	0	0	0.090909	0.014621
SARs column constraint	63.0541	9.1133	0.985222	13.38259	13.4647	Σ = 100

Table 11 – Using IPF to spatially disaggregate the SARs probability distributions

In particular, this table refers to the employment status of heads of households who are *male*, *single-widowed-divorced* and aged between *16-19 years*. The column constraints of this table are derived from the SARs and depict the employment status percentages of the above population group for the whole Leeds Metropolitan District. The row constraints of the table are derived from the SAS and depict the percentages of this population group in each ED of Leeds. The cells of the matrix inside the table

contain the employment status probability distribution of this population group within each ED provided by the SAS. This kind of information can also be seen as 'interaction' data because it gives the *relative weight* of each cell. IPF can be applied on this matrix so that its row and column totals are equal to the row and column constraints respectively. In this manner, we can derive a probability distribution for the following household characteristics:

- location (ED level)
- age of head of household
- sex of head of household
- marital status of head of household
- employment status of head of household

Figure 2 shows the estimated rates for employees in Leeds by Enumeration District, while figure 3 provides the actual rates, which are derived from the SAS. In addition, figure 4 depicts the absolute difference between estimated and actual rates.

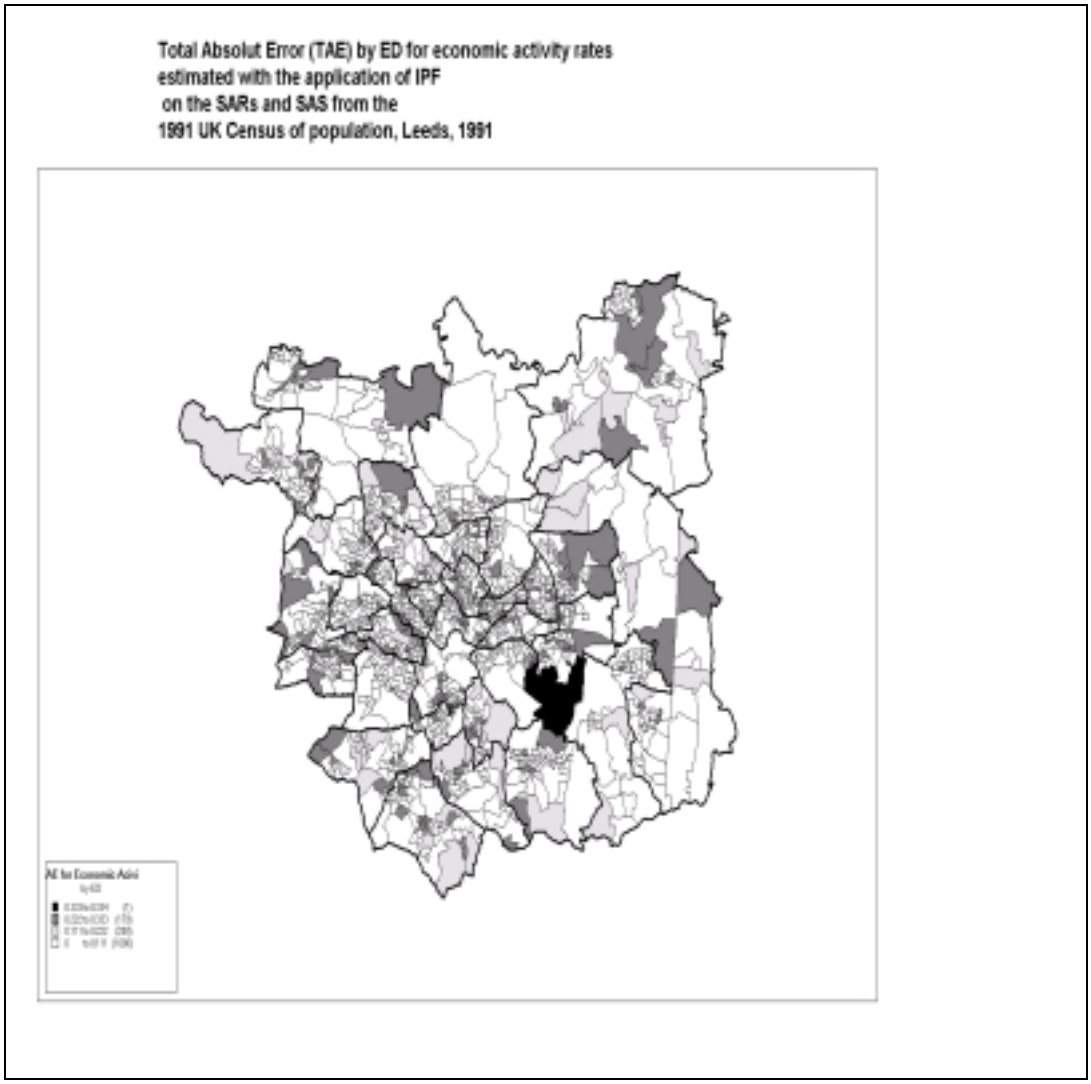


Figure 4

It should be noted that the IPF procedure could be used to estimate the distribution of conditional probabilities for more household attributes. Figures 5 and 6 provide examples of IPF applications for the estimation of conditional probabilities for more household attributes.

4.4.2 SimLeeds3: Monte Carlo sampling from SAS generated conditional probabilities with the use of IPF

This model uses conditional probabilities, which are derived, with the use of the IPF technique on different SAS tables, as illustrated in section 3.4.2. Model *SimLeeds3* follows a bottom-up micro-modelling approach, as it was the case with the models described above. In particular, it requires a baseline population of households and then uses conditional probabilities derived from different SAS tables with the use

of IPF in order to assign attributes to the baseline household population. The classes and class variables of *SimLeeds3* are the same with those of the previous two models. Moreover, the steps followed by *SimLeeds3* are also similar to those of the previous models and they are as follows:

- **Step 1** - Read the baseline household population and the probability files (generated from different SAS tables with the use of IPF)
- **Step 2** - Create an array of household objects based on the baseline population data
- **Step 3** - Assign each household the characteristics of table 12 on the basis of random (Monte Carlo) sampling, using the conditional probability distributions given in the probability files.
- **Step 4** - Output the list of microsimulated households

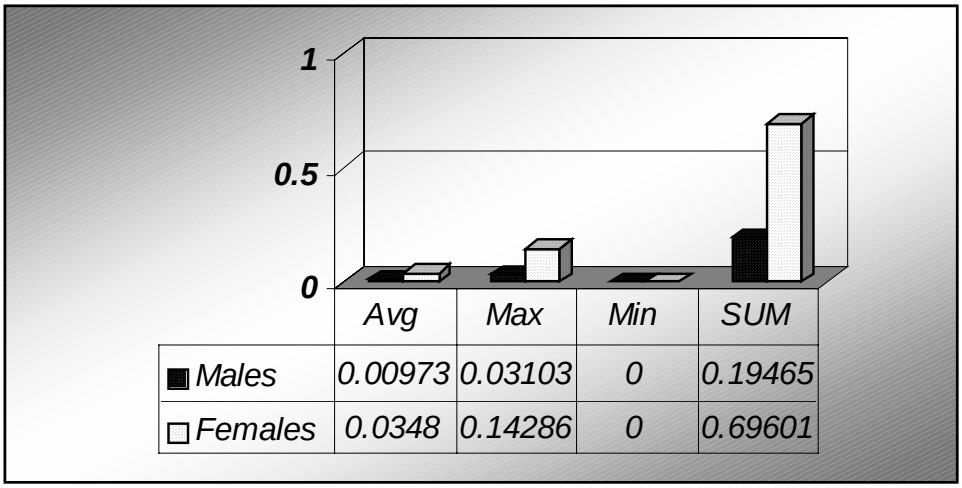
Tables 25 and 26 present the TAE of SimLeeds3 for males and females and graph 4 depicts the TAE statistics.

Age	Employee	Self-employed	On a Government scheme	Unemployed
16-29	0.005	0.013	0.009	0.017
30-44	0.031	0.006	0.019	0.007
45-59	0.010	0.007	0.002	0.015
60-64	0.028	0.000	0.000	0.028
65+	0.000	0.000	0.000	0.000

Table 25 - Employment status TAE – absolute difference between estimated and observed probability rates of males, ED 1, Beeston, Leeds, SimLeeds3 (five-run average)

Age	Employee	Self-employed	On a Government scheme	Unemployed
16-29	0.022	0.004	0.004	0.014
30-44	0.081	0.000	0.000	0.081
45-59	0.060	0.021	0.021	0.102
60-64	0.143	0.000	0.000	0.143
65+	0.000	0.000	0.000	0.000

Table 26 - Employment status TAE – absolute difference between estimated and observed probability rates of females, ED 1, Beeston, Leeds, SimLeeds3 (five-run average)



	Avg	Max	Min	SUM
Males	0.00973	0.03103	0	0.19465
Females	0.0348	0.14286	0	0.69601

Graph 5 - Employment status – TAE statistics for males and females, SimLeeds3 (five-run average)

As can be seen, SimLeeds3 error statistics are slightly better than those of *SimLeeds2* and *SimLeeds1* are.

4.4.3 SimLeeds4: Monte Carlo sampling from SARs and SAS generated conditional probabilities with the use of IPF

This model is designed in slightly different way from the previous three. In particular, *SimLeeds4* follows a top-down micro-modelling approach and performs random sampling from SARs and SAS generated conditional probabilities with the use of IPF, as illustrated in section 4.4.1.3. Further, as in the previous modelling approaches, there are two classes: *SimLeeds* and *Household*. Nevertheless, the Household class is different in that it only has one instance variable depicting the Household type. Under this modelling approach there are five basic steps:

Step 1 – Get conditional probabilities from the SARs and the SAS related to the household attributes that need to be estimated (in a way similar to that depicted in table 11).

Step 2 – Apply the Iterative Proportional Fitting technique to the above probabilities in order to generate a matrix of micro-probability distribution.

Step 3 – Generate a list of Household objects ('instantiate' the Household class)

Step 4 – Read the above micro-probability matrix and assign type to each Household on the basis of random sampling

Step 5 – Export the list of the estimated Households with their associated attributes.

Step 6 – Export a table with the counts of Household of each different category, which can be easily integrated with spatial boundary data in a GIS.

An illustrative example is useful at this point to clarify the modelling procedure. From the 1991 Census of the UK population (SAS table 8) we can obtain the employment status rates by broad age-group and sex at the ED level. Moreover, from the Samples of Anonymised Records we can obtain conditional probabilities for additional characteristics of the above population group, but with the price of losing in geographical detail (the finest geographical scale at the individual level is the Metropolitan District). Nevertheless, the above data sets can be integrated with the use of IPF. In particular, IPF can be used to spatially distribute the SARs probabilities to all the Leeds EDs. For instance, table 24 depicts the estimated (with the use of IPF) conditional probabilities of 16-29 year olds in Leeds with respect to the following attributes:

- Location at the ED level (residence) (1,429 possible states)
- Employment status (5 possible states)
- Marital status (2 possible states)
- Tenure (3 possible states)
- Educational qualifications (4 possible states)

Location	Atr →	1	2	3	4	5	6	7	8	9	.	.	.	120
DAFA01	1	0.006	0.001	0.000	0.001	0.006	0.000	0.001	0.000	0.000	.	.	.	0.000
DAFA02	2	0.020	0.004	0.000	0.001	0.009	0.000	0.002	0.000	0.000	.	.	.	0.000
DAFA03	3	0.022	0.002	0.000	0.001	0.006	0.000	0.001	0.000	0.000	.	.	.	0.000
DAFA04	4	0.024	0.001	0.000	0.002	0.007	0.000	0.001	0.000	0.000	.	.	.	0.000
DAFA05	5	0.011	0.001	0.000	0.001	0.007	0.000	0.000	0.000	0.000	.	.	.	0.000
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
DAGK45	1429	0.021	0.003	0.000	0.000	0.003	0.000	0.001	0.000	0.000	.	.	.	0.000

Table 24 – Micro-probability distribution matrix, generated with the use of IPF

The employment status rates at the ED level were obtained from the SAS table 8. The following step was to use the USAR package[3] in order to retrieve data for the additional characteristics (i.e. marital status, tenure, educational qualifications) using the attributes included in SAS table 8 as filters. Then, IPF was applied to 'spatially distribute' the SARs probabilities to the Eds. As can be seen in table 24, there are 1429 spatial zones (enumeration districts) and 5 x 3 x 4 x 2 = 200 possible household socio-economic states. Thus, there are 1429 x 5 x 3 x 4 x 2 = 171,480 states or household types. For example, the probability depicted at the above *matrix element (1,1)* represents the probability of someone being a single-widowed-divorced (SWD)

household head, aged 16-29 years, living in Leeds ED DAFA01, having no educational qualifications and being an employee and owner occupied. *SimLeeds4* reads the above matrix and performs random sampling. It should be noted that *SimLeeds4* requires a given number of households to simulate (baseline household population). In this case, the household constraint can be the total number of 16-29 Single-Widowed-Divorced males in Leeds, which are 71,487. In particular, from the 1991 Census of the UK population we can obtain the employment status rates by age-group and sex at the ED level, and in this case for the 16-29 year old males. Moreover, from the Samples of Anonymised Records we can obtain additional characteristics for the above population group, but we would loose the geographical detail. IPF can be used on the above data sets (from the SAS and the SARs) to spatially distribute the SARs population to all the Leeds EDs. *SimLeeds4* returns a list of 71,487 households with their associated types.

Figures 5 and 6 depict the geographical distribution of two estimated microgroups using *SimLeeds4*. In particular, figure 5 represents the estimated spatial distribution at the ED level of males, single or widowed or divorced (SWD) who are aged 16-29 and economically active, owner occupied, without formal educational qualifications and live in the Leeds Metropolitan District. According to the model, the size of this micro-group of the Leeds population is 18,620 individuals. This estimate represents the 26% of the total Leeds male population aged between 16-29, whereas the actual respective percentage provided by the SARs is 25%. Therefore, there is no large deviation between the actual (observed) and the estimated (microsimulated) percentage at this level. Likewise, figure 6 depicts the estimated geographical distribution at the ED level of SWD unemployed males who live in rented accommodation and have no formal educational qualifications. The estimated size of this micro-group is 4,667 individuals, which represents the 6.52% of the total Leeds male population aged between 16-29, whereas the actual percentage provided by the SARs is 6.59%.

[3] For more information on the USAR see Turton and Openshaw, 1994a; Middleton, 1995.

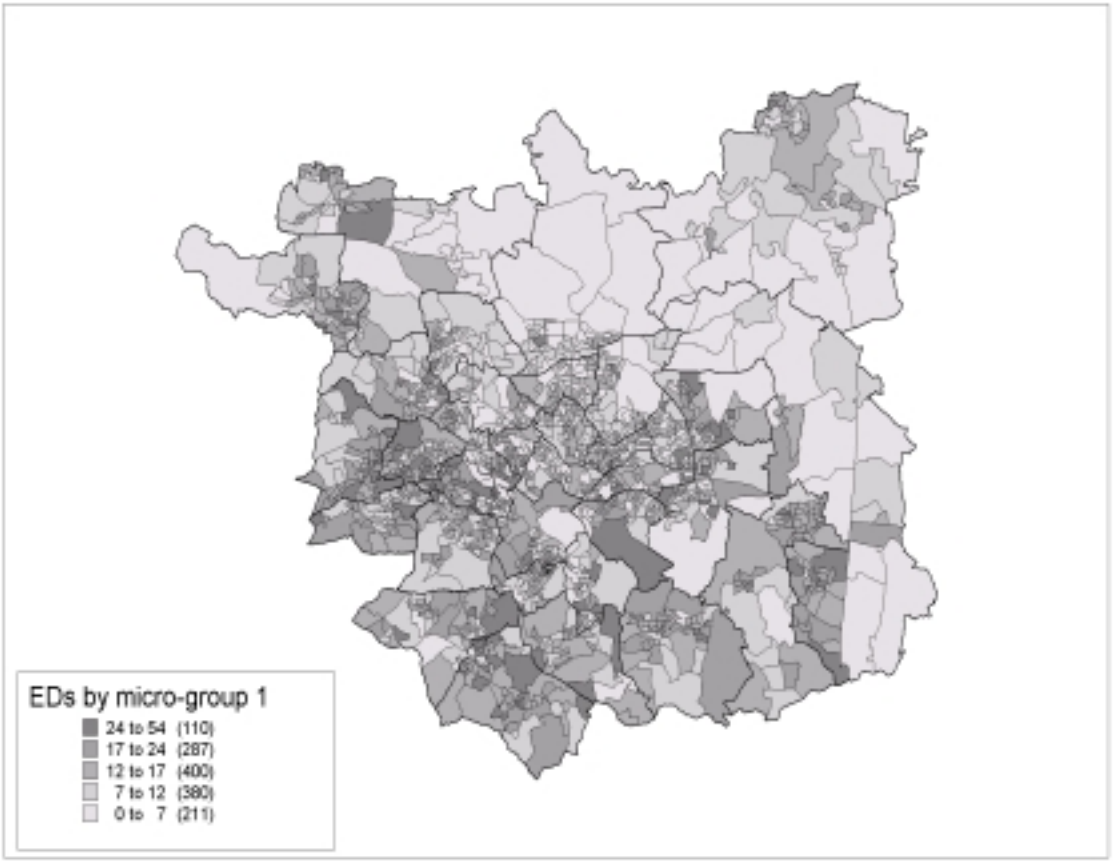


Figure 5 – Spatial distribution of Micro-group 1: males, SWD aged 16-29, economically active, employees, owner occupied, without formal educational qualifications, *SimLeeds4*.

As it was the case with the other models, it is quite difficult to validate the model outputs at the ED level, since there are no available microdata at this geographical scale. However, one way of validating the analysis is to extract distributions from the estimated microdata, for which observed data exist (e.g. produce tables comparable to published Census cross-tabulations). For instance, figure 7 depicts the spatial distribution of the Total Absolute Error for unemployed, male Leeds residents who are aged 16-29. More specifically, this figure shows the spatial variations of the difference between the observed and the estimated distributions of this population group. As can be seen, there are 17 EDs with a relatively high TAE, ranging between 0.159 and 0.6, whereas the majority of the EDs (746) have a very small TAE (ranging between 0 and 0.033). Likewise, figure 8 shows the spatial distribution of the TAE for male employees, who are Leeds residents and are aged 16-29. As can be observed in this map, there are generally higher values of TAE for this population group. However, the majority of EDs (1318 out of 1429) have a TAE, which is less than 0.211.

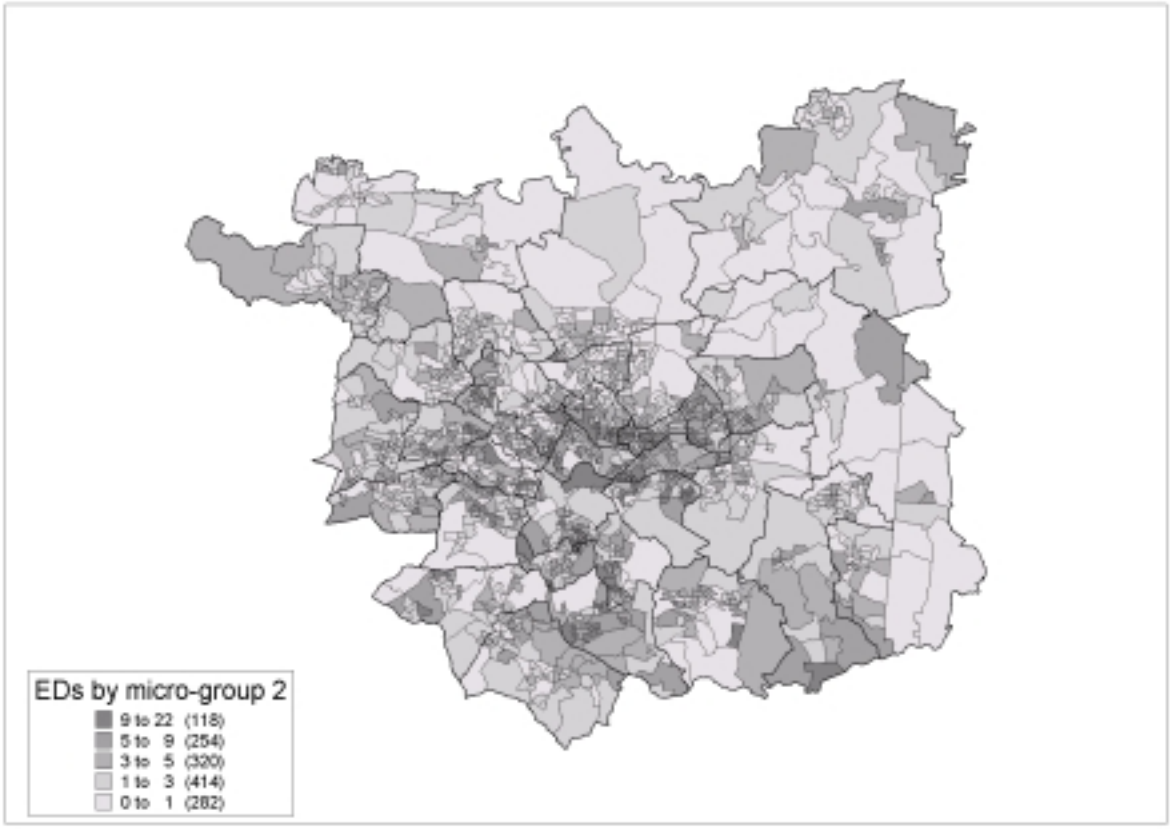


Figure 6 – Spatial distribution of Micro-group 2: SWD unemployed males, rented accommodation, no formal educational qualifications, *SimLeeds4.*

It should be noted that the big advantage of this modelling approach is that it can easily generate estimates of microsimulated population groups at the ED scale, for all the desired EDs, whereas the previous bottom-up approaches require to put more effort on model designing in order to achieve this.

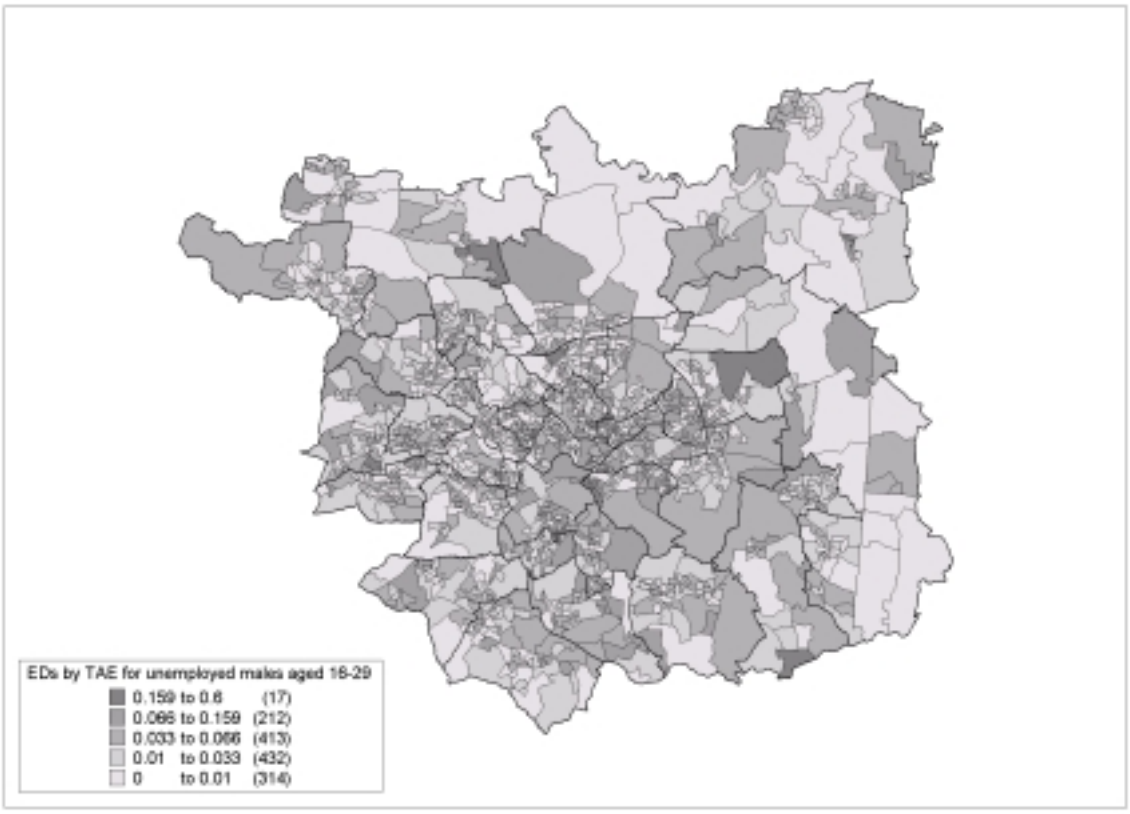


Figure 7 - Spatial distribution of TAE for unemployed, male Leeds residents, aged 16-29, *SimLeeds4*

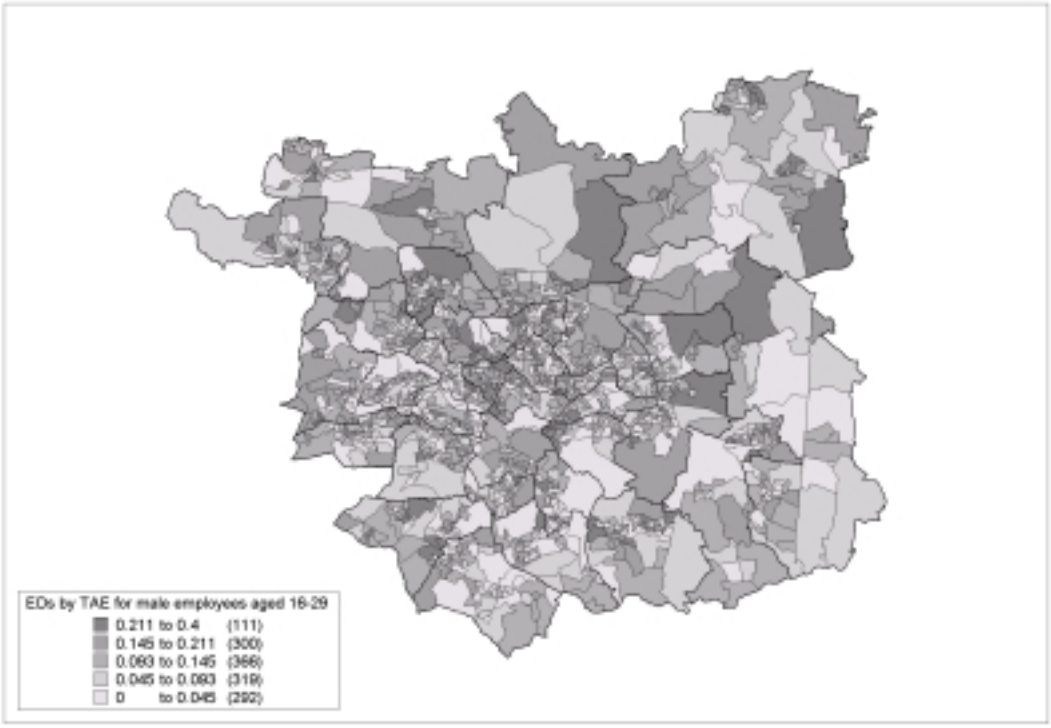


Figure 8 - Spatial distribution of TAE for male employees, Leeds residents, aged 16-29, *SimLeeds4*

4.5 SimLeeds5: Reweighting a microdata sample to estimate spatially disaggregated microdata

4.5.1 Reweighting a microdata sample

This section describes an alternative way of generating a population that matches the constraints for a small area provided by SAS tables. As mentioned in section 4.3.1 the Sample of Anonymised Records (SARs) provide a detailed record for a two percent sample of British households and all of their occupants. These records are only geographically referenced to the standard region level thus making them less useful for studies of small areas. The reweighting method described here samples this universe of records to find the set of household records that best matches the population described in the SAS tables for the small area under study.

The method works by considering a series of SAS tables that describe the area and selecting records, which are then added to the tables being constructed. There are a vast number of possible sets of approximately 200 households that can be drawn from the sample of 541,894 records in the whole sample. Clearly, it would be impractical to exhaustively consider all possible sets so this study uses simulated annealing (Dowsland, 1993) to find a set that fits the target tables well. An initial random sample of records is selected until sufficient households are represented. These records are then used to create tables that match the selected target tables. Each pair of tables is then compared to calculate the absolute error between the two tables. A record in the set is then selected at random and replaced with one chosen at random from the universe of records. The error is then recalculated and the change in error (Δe) is calculated. If Δe is less the zero then there has been an improvement and the change is accepted, if not then $\exp(-\Delta e/t)$ is compared to a random number between 0 and 1, if it is greater than the random number then the change is accepted, else the change is rejected and reversed. The method works as an analogy with the physical process of cooling a metal slowly to allow larger crystals to form. A temperature is initially set high and then slowly lowered after a set number of iterations have taken place. As the temperature is lowered fewer uphill moves are accepted. This allows the method to recover from being stuck in local minimum by allowing a limited number of small moves uphill. In this implementation if Δe is zero the change is accepted to allow the exploration of a greater part of the solution space. If the new error is the best

seen so far the set of households used is stored. This procedure is summarised in figure 9.

This method has one key advantage over the other methods presented in this paper which is that the population produced are “real” people living in “real “ households. This allows the user to create new tables that are either not included in the SAS or to produce tables that are provided in the SAS but without the need to rerun the modelling exercise.

There are some specific problems associated with this process of selecting individual household records to fill SAS tables. The first one is that to protect the confidentiality of the Census the OPCS applied a blurring process to the tables before they were released. Each cell in a table was modified by –1, 0 or +1 in a quasi-random pattern. To avoid negative numbers in tables no modification was applied to precise zeros in a table. This modification procedure was applied in a probabilistic manner such that there is a less than 5 percent chance that the sum of ten cells differs by more than +/-3 from the actual count (Marsh, 1993). One unfortunate result of this process is that two tables for the same area are unlikely to sum to the same total, that is there appear to be different numbers of people and households in a given area depending on which table is studied. This leads to the problem that the simulated annealing method is very unlikely to ever reach an absolute error of zero and will always run until the iteration limit is exceeded.

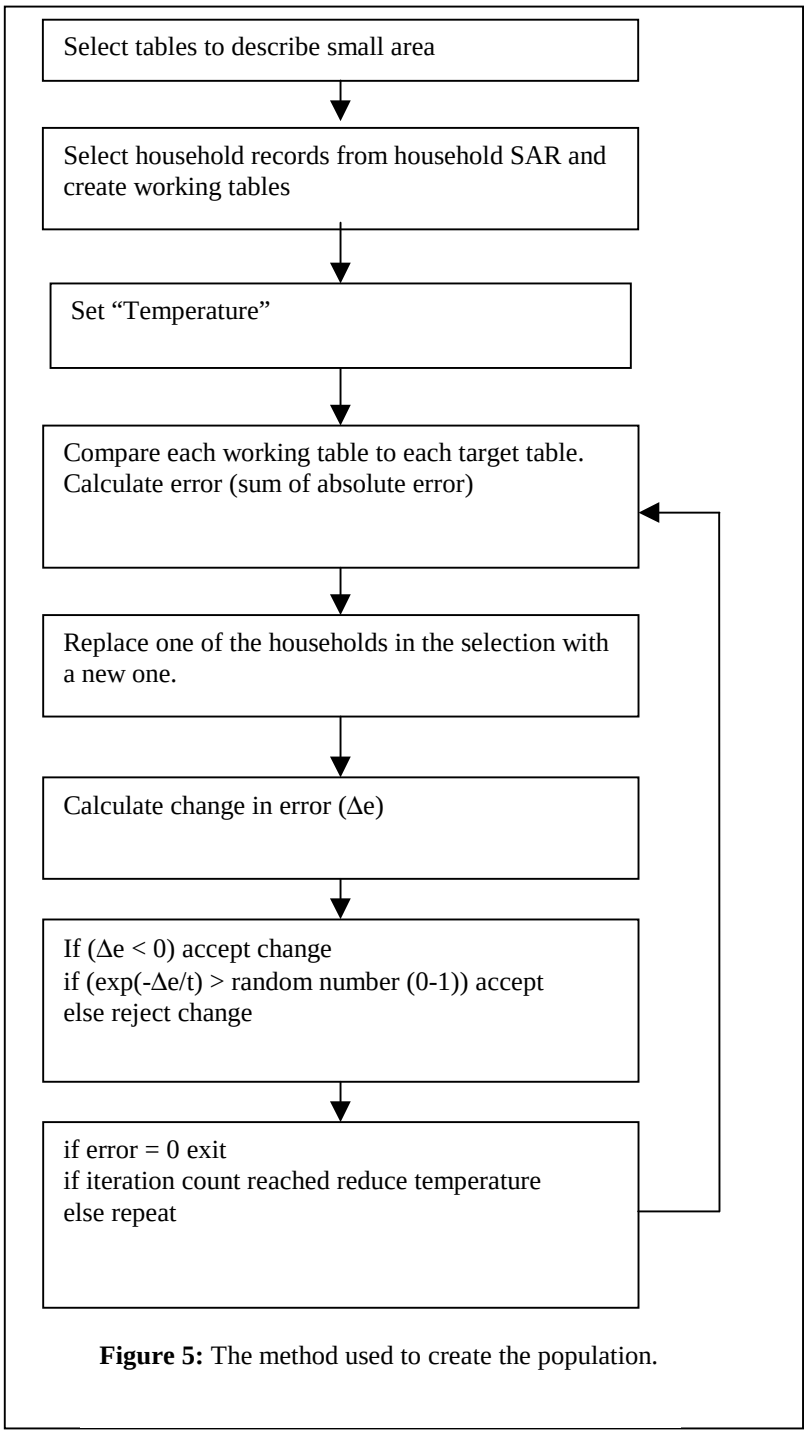


Figure 5: The method used to create the population.

The second problem that can be encountered is the difficulty of matching variables that are used to construct tables with variables that are available in the SARs. In general the variables used in each output are derived from the same questions, however there are some variables that have been grouped or top coded in both sets to preserve data confidentiality.

4.5.2 SimLeeds5: Reweighting the SARs

This model uses a similar class structure to the previous models to store households and the individuals within them. The structure of the program is however different and the algorithmic steps are:

- **Step 1** – read in SAS tables (raw counts not probabilities).
- **Step 2** – read in SAR database in to household records.
- **Step 3** – select sufficient households at random to populate the tables.
- **Step 4** – apply simulated annealing to find best fitting set of households.
- **Step 5** – when error = 0 or iteration count exceeded write out best set of records.

The SAS tables chosen to constrain the samples for this study were as follows:

- Table S06 Sex by Ethnic Group
- Table S08 Economic Activity by Age
- Table S35 Age by Marital Status
- Table S39 Age of Head of Household by Marital Status.

There is a trade off between more tables giving a better population match to reality and more tables adding to the computational burden of the program. It was felt that four tables would be sufficient to provide a reasonable fit to the underlying population without making the optimisation problem so hard that it became insoluble.

In this case it is possible to consider the comparison of the tables produced with published tables (while bearing in mind the comments made about data blurring mentioned above).

These tables allow us to see how by using full households with individual records inside them that it is possible to consider both individual tables and tables based on the head of household. In all of the tables used the error per cell is between 0.3 and 0.5 which can be considered reasonable.

S06 Sex by Ethnic group

	White	Black Caribb	Black African	Black Other	Indian	Pakistani	Bangladeshi	Chinese	Other-Asian	Other	Total
Male	231	0	0	1	3	1	0	0	1	0	237
Female	197	0	0	0	0	2	0	0	1	0	200
Total	428	0	0	1	3	3	0	0	2	0	437

S06 Sex by Ethnic group (estimates)

	White	Black Caribb	Black African	Black Other	Indian	Pakistani	Bangladeshi	Chinese	Other-Asian	Other	Total
Male	229	2	0	0	2	1	0	0	2	1	237
Female	196	0	0	1	0	2	0	0	1	0	200
Total	425	2	0	1	2	3	0	0	3	1	437

S06 Sex by Ethnic group (error)

	White	Black Caribb	Black African	Black Other	Indian	Pakistani	Bangladeshi	Chinese	Other-Asian	Other	Total
Male	2	2	0	1	1	0	0	0	1	1	8
Female	1	0	0	1	0	0	0	0	0	0	2
Total	3	2	0	2	1	0	0	0	1	1	10

Error per cell = 0.5

S39 HOH age by marital status

	Male SWD	Male Mar	Female SWD	Female Mar	Total
g2	33	7	13	2	55
g3	34	16	14	0	64
g4	15	19	18	3	55
g5	7	9	3	0	19
g6	5	9	9	0	23
g7	3	4	9	1	17
g8	0	1	1	0	2
Total	97	65	67	6	235

S39 HOH age by marital status (estimated)

	Male SWD	Male Mar	Female SWD	Female Mar	Total
g2	33	7	13	1	54
g3	34	17	15	3	69
g4	14	19	16	3	52
g5	7	9	3	0	19
g6	5	9	9	0	23
g7	3	4	9	1	17
g8	0	1	0	0	1
Total	96	66	65	8	235

S39 HOH age by marital status (error)

	Male SWD	Male Mar	Female SWD	Female Mar	Total
g2	0	0	0	1	1
g3	0	1	1	3	5
g4	1	0	2	0	3
g5	0	0	0	0	0
g6	0	0	0	0	0
g7	0	0	0	0	0
g8	0	0	1	0	1
Total	1	1	4	4	10

Error per cell = 0.35

5. Concluding comments

In this paper we have explored different spatial microsimulation approaches for the estimation of household attributes. We have outlined the merits and drawbacks of these approaches and we have presented 5 spatial microsimulation models for Leeds. It can be reasonably argued that the models *SimLeeds4* and *SimLeeds5* are the most robust and powerful, although, as seen above, it is quite difficult to validate microsimulation model outputs. *SimLeeds4* has the advantage of generating spatially disaggregated microdata for a city or region relatively fast. It is also relatively easy to generate the probabilities in a suitable form that is required from the model. On the other hand, *SimLeeds5*, which represents a simulated annealing approach to microsimulation modelling has the added advantage of generating "real" people living in "real " households. It should be noted that both *SimLeeds4* and *SimLeeds5* are computationally intensive and they require further work on the development of data input procedures, in order to simulate the population of entire cities or regions more efficiently.

Our future work will thoroughly test *SimLeeds4* and *SimLeeds5* in order to have a more informed decision on which one is the most reliable. Further, other possible combinations of modelling approaches will be explored. Our ultimate goal is to use our *SimLeeds* series of spatial microsimulation models in order to perform *what-if* geographical analysis and to use them as tools for the evaluation of regional and urban policies at the micro-scale.

Acknowledgements

The work reported on that paper was part funded by the Greek State Scholarships Foundation (IKY). The Census Small Area Statistics are provided through the Census Dissemination Unit of the University of Manchester, with the support of the ESRC / JISC / DENI 1991 Census of Population Programme. The Census Sample of Anonymised Records are provided through the Census Microdata Unit of the University of Manchester, with the support of the ESRC / JISC / DENI. All Census data reported in this paper are Crown Copyright. The authors would like to thank Paul Norman for providing his program IPFprog.

References

Batty, M. (1996) *Review of C.Bertuglia, G.P.Clarke, A.G.Wilson (eds) Modelling the city*, Routledge, London, in Progress in Human Geography, *20(2*), 260-262

Betson D., Greenberg D., Kasten R. (1980) *A microsimulation model for analysing alternative welfare reform programmes for better jobs and income*, in

Microsimulation models for public policy and analysis, vol 1, Academic Press, New York.

Birkin, M. (1987), *Iterative Proportional Fitting (IPF): theory, method, and example*, Computer Manual 26, School of Geography, University of Leeds, Leeds

Birkin M. (1995) *Customer targeting, geodemographics and lifestyle approaches,* in P.Longley and G.P.Clarke (eds) GIS for business and service planning, Geoinformation, Cambridge, 104-149

Birkin, M. , Clarke, G. (1995), *Using microsimulation methods to synthesize census data*, in S. Openshaw (ed.), Census Users' Handbook, GeoInformation International, London, pp. 363-387

Birkin, M. , Clarke, G.P. , Clarke, M. (1996), *Urban and regional modelling at the microscale*, in G.P. Clarke (ed.) , Microsimulation for Urban and Regional Policy Analysis, Pion, London.

Birkin M., Clarke 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., Clarke M. (1989) *The generation of individual and household incomes at the small area level*, **Regional Studies**, *23(6),* 535-548

Caldwell S. (1986) *Broadening policy models: alternative strategies*, in G.H.Orcutt, J.Merz and H.Quinke (eds) Microanalytic simulation models to support social and financial policy, North-Holland, Amsterdam, 59-76

Caldwell S., Keister L. (1996*) Simulating the accumulation and distribution of corporate stock holdings by US families*, in G.P.Clarke (ed) *Microsimulation for urban and regional analysis*, Pion, London.

Clarke, G. P. (1996) , *Microsimulation: an introduction*, in G.P. Clarke (ed.) , Microsimulation for Urban and Regional Policy Analysis, Pion, London.

Clarke M. (1986) *Demographic processes and household dynamics: a microsimulation approach*, in R.Woods and P.H.Rees (eds,*)* Population structures and models: developments in spatial demography, Allan & Unwin, London, 245-272

Clarke M., Holm E. (1987) *Micro-simulation methods in human geography and planning: a review and further extensions*, Geografiska Annaler, *69B*, 145-164

Duley C.J., Rees P.H., Clarke M. (1988) *A microsimulation model for updating households in small areas between Censuses*, Working paper 515, School of Geography, University of Leeds.

Dowsland, K., (1993), *Simulated Annealing*, in Reeves, C. (ed) Modern Heuristic Techniques for Combinatorial Problems, Blackwell, Oxford.

Edwards J. (1995) *Social policy and the city: a review of recent policy developments and literature*, Urban Studies, *32*, 695-712

Dowsland, K., (1993), *Simulated Annealing*, in Reeves, C. (ed) Modern Heuristic Techniques for Combinatorial Problems, Blackwell, Oxford.

Fienberg, S.E., (1970), *An iterative procedure for estimation in contingency tables*, Annals of Mathematical Statistics **41**, 907-917

Flanagan, D. (1997), *Java in a Nutshell***,** O'Reilly, Sebastopol

Grand, M. , Knudsen, J. (1997), *Java Fundamental Classes Reference*, O'Reilly, Sebastopol

Green A. (1994) *The geography of poverty and wealth: evidence of the changing spatial distribution and segregation of poverty and wealth from the Census of Population 1991 and 1981*, Institute for Employment Research, University of Warwick.

Hill D. (1994) *Citizens and cities: urban policy in the 1990s*, Harvester Wheatsheaf, London.

Hills J. (1993) *The future of welfare: a guide to the debate*, Joseph Rowntree Foundation, York.

Holm E. et al (1996) *Full-scale microsimulation: a tool for policy analysis*, in G.P.Clarke (ed) Microsimulation for urban and regional analysis, Pion, London.

Johnson N. (1990) *Reconstructing the welfare state: a decade of change 1980-1990*, Harvester Wheatsheaf, London.

Kain S., Apgar W. (1985) *Housing and neighbourhood dynamics: a simulation study*, Harvard University Press, Cambridge.

Lemay, L., Perkins, C.L. (1997), *Teach yourself Java 1.1 in 21 days*, second edition, SamsNet, Indianapolis

Marsh, C (1993), *The sample of anonymised records*, in A Dale and C Marsh (eds), The 1991 Census Users's Guide, London, HMSO, 295-311

Marsh, C., (1993), *Privacy, Confidentiality and anonymity in the 1991 Census*, in Dale, A and Marsh, C. (eds) The 1991 Census User's Guide, HMSO, London.

Marsh, C and Teague A (1992), *Samples of anonymised records from the 1991 Census*, Population Trends, **69**, pp. 17-26.

Mertz, J. (1991), *Microsimulation - A survey of principles developments and applications*, International Journal of Forecasting, **7**, pp.77-104

Middleton, E. (1995), *Samples of Anonymized Records***,** in S. Openshaw (ed.), Census Users' Handbook, GeoInformation International, London, pp. 337-362

Noble M., Smith G., Aveneu D., Smith T. and Sharland E. (1994*) Changing patterns of income and wealth in Oxford and Oldham*, Department of Applied Social Studies and Social Research, University of Oxford.

Norman, P. (1999), *Putting Iterative Proportional Fitting on the Researcher's Desk*, forthcoming working paper, School of Geography, University of Leeds, Leeds

Openshaw S. (1995a*) Human systems modelling as a new grand challenge area in science*, Environment and Planning A, 27(2), 159-164

Openshaw, S. (ed.) (1995), *Census Users' Handbook*, GeoInformation International, London.

Orcutt, G.H., Mertz J., Quinke, H (Eds.), (1986), *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam

Rees, P. (1994) *Estimating and Projecting the Populations of Urban Communities*, Environment & Planning A**, 25**, pp. 1671-1697

Rees P.H., Clarke M., Duley C. (1987*) A model for updating individual and household populations*, Working paper 486, School of Geography, University of Leeds.

Thornley A. (1991) *Urban planning under Thatcher*, Routledge, Andover.

Wegener M. (1986) *Integrated forecasting models or urban and regional systems*, in P.J.Batey and M.Madden (eds) Integrated analysis of regional systems, Pion, London.

Williams H.C.W., Keys P., Clarke M. (1986*) Vacancy chain models for housing and employment systems*, Environment and Planning A, **18**, 89-105

Williamson P., Birkin M., Rees P.H. (1993) *The simulation of whole populations using data from small area statistics, samples of anonymised records and national surveys*, Paper presented at the 'Research on 1991 Census conference', University of Newcastle, September

Turton, I. , Openshaw, S. (1994a), *A step-by-step guide to accessing the 1991 SAR via USAR*, Working Paper-94/6,School of Geography, University of Leeds.

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

Williamson, P., Clarke, G.P., McDonald, A.T. (1996), *Estimating small area demands for water with the use of microsimulation*, in G.P. Clarke (ed.), Microsimulation for Urban and Regional Policy Analysis, Pion, London.

Williamson, P., Birkin, M., 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**, pp. 785-816

Wong, D W S (1992), *The Reliability of Using the Iterative Proportional Fitting Procedure*, Professional Geographer, **44**, 1992, pp. 340-348