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

SimBritain: A Spatial Microsimulation Approach to Population Dynamics

Dimitris Ballas,[1]* Graham Clarke,[2] Danny Dorling,[1] Heather Eyre,[3] Bethan Thomas[1] and David Rossiter[2]

[1] *Department of Geography, University of Sheffield, Winter Street, Sheffield S10 2TN, UK*
[2] *School of Geography, University of Leeds, Leeds LS2 9JT, UK*
[3] *Education Leeds, Merrion House, 110 Merrion Centre, Leeds LS2 8DR, UK*

ABSTRACT

In this paper we present an account of a 3-year research project that is aimed at dynamically simulating urban and regional populations in Britain. In the context of this project we are using data from the 1991 UK Census Small Area Statistics (SAS) and the British Household Panel Survey (BHPS), in order to dynamically simulate the entire population of Britain into 2021 at the small area level. This paper discusses the structure, aims and objectives of *SimBritain* and presents some preliminary results. Firstly, alternative spatial microsimulation strategies are discussed and their advantages and drawbacks are outlined. Next, the difficulties in calibrating and validating dynamic microsimulation models such as *SimBritain* are highlighted and ways to tackle these difficulties are explored. The paper then presents some model outputs that highlight the geographical variation of a wide range of socio-economic variables through the 1990s. Moreover, in light of these outputs, the paper discusses the potential of *SimBritain* for policy analysis.

INTRODUCTION

Microsimulation models are concerned with the creation of large-scale datasets estimating the attributes of individuals within households. Further, microsimulation techniques involve the use of these attributes for policy analysis at the micro-level. In particular, by permitting analyses at the level of the individual and household, they provide the means of assessing variations in the distributional effects of different policies (Mertz, 1991; Hancock and Sutherland, 1992). In addition, microsimulation modelling frameworks provide the possibility of estimating whether the goals of economic and social policy might be achieved: where they might be achieved and where not (Krupp, 1986). Microsimulation methodologies are becoming accepted tools in the evaluation of economic and social policy, in the analysis of tax-benefit options, and in other areas of public policy (Hancock and Sutherland, 1992).

It can be argued that there is still very little use of microsimulation methodologies within geography and regional science that aims to dynamically simulate populations at the small area level. In this paper we explore spatial dynamic microsimulation approaches. We also present *SimBritain*, which is a spatial dynamic microsimulation model that utilises the British Household Panel Survey (BHPS) in order to simulate the population of Britain under different scenarios. The ultimate aim is to develop a policy-relevant spatial microsimulation model that would be used for socio-economic impact assessment. *SimBritain* aims at simulating the entire population of Britain at the ward level. In order to render this task more manageable we have developed *SimYork*, which is a model of the future population of the city of York.

SimYork is being used as a pilot study in order to test different methodologies and combinations of datasets. In this paper we present results of *SimBritain* at the geographical level of parliamentary constituencies. We also present some model outputs of *SimYork* at the electoral ward level which highlight the key issues in simulating small area populations. It should be noted that the main difference between *SimYork* and *SimBritain* is that the former was developed and implemented at the electoral ward level, whereas the latter was implemented at the parliamentary constituency level. Another difference of a more technical nature between the two modelling approaches is that the former has been developed in the Java programming language,[1] whereas the latter has been developed and implemented in SAS.[2]

The aim of this paper is to present the modelling methodology that we adopted and to show some preliminary results. Our overall aim is to use the modelling approach presented here in order to perform policy analysis (for an example of this kind of analysis, see Ballas *et al*., 2004). However, it should be noted that this paper focuses on the more technical aspects of the research. Firstly, we briefly review past attempts to implement microsimulation methodologies on populations. We then introduce the *SimBritain* model and we discuss in some detail the methodologies employed to simulate Census small area statistics tables into the future. In particular, we describe the data-sets that we used and we discuss the modelling methodology that we developed and employed to project small area statistics tables into the future. We present a reweighting technique that we developed and used to readjust the weights of the BHPS records, so that they would fit small area statistics tables, and show how the *SimBritain* model can be validated. Finally, some concluding comments are offered.

SPATIAL MICROSIMULATION APPROACHES TO INDIVIDUAL AND HOUSEHOLD DYNAMICS

During the 1990s many studies of British society made use of some form of microsimulation (e.g. Dorling and Woodward, 1996; Williamson, 1996; Brimblecombe *et al*., 2000; Ballas and Clarke, 2001a). The technique has grown in use recently as the computer-intensive nature of the work has become less of a handicap. However, it can be argued that some of these studies did not recognise that the techniques they were using fell into this category and they often had to 're-invent the wheel' in their work.

Microsimulation models can be categorised into various types. For instance, Mertz (1991) classified them as either *static models*, based on simple snapshots of the current circumstances of a sample of the population at any one time, and *dynamic models* that vary or age the attributes of each micro-unit in a sample to build up a synthetic longitudinal database describing the sample members' lifetimes into the future. Falkingham and Lessof (1992) provided a more detailed categorisation of the methods. The work presented in this paper aims at utilising both of these main modelling approaches. The microsimulation method typically involves four major procedures:

(1) The construction of a microdata-set from samples and surveys (when microdata are not available from published sources).
(2) Sampling from this dataset to 'create' a micro-level population for individuals for small areas who match the known data on those areas.
(3) *Static What-if* simulations, in which the impacts of alternative policy scenarios on the population are estimated: for instance, if there had been no poll tax in 1991, which communities would have benefited most and which would have had to have paid more tax in other forms?
(4) Dynamic modelling, to update a basic microdata- set and create future-oriented *what-if* simulations: for instance, if the current government had raised income taxes in 1997, what would the redistributive effects have been between different socio-economic groups and between central cities and their suburbs by 2011?

It has long been argued that, in most cases, the microsimulation models developed so far do not take spatial scale into account (Birkin *et al*., 1996). Among the notable exceptions has been the work conducted at Leeds University (Birkin and Clarke, 1988) which used static spatial microsimulation techniques to generate a synthetic microdatabase for the Leeds Metropolitan District.

Moreover, Birkin and Clarke (1989) used this SYNTHESIS model to generate incomes for individuals. Williamson (1992) developed a model for the spatial analysis of community care policies for the elderly, whereas Ballas *et al*. (1999) and Ballas and Clarke (2000) argued and demonstrated the case for a geographical microsimulation approach to local labour-market analysis and developed a spatial microsimulation model for the Leeds local labour market. In addition, Ballas and Clarke (2000, 2001a) adopted a spatial microsimulation approach to the analysis of the spatial impacts of social policies. In particular, they used spatial microsimulation methodologies to perform static *what-if* policy analysis and identify which types of households and which geographical areas would be affected the most under different social policy scenarios. In another study (Ballas and Clarke, 2001b), static spatial microsimulation was used to evaluate the socioeconomic impact of a plant closure in East Leeds. Another example of spatial microsimulation modelling is the work of the Spatial Modelling Centre (SMC)[3] in Sweden (Holm *et al.*, 1996; Vencatasawmy *et al.*, 1999) where researchers have built a geographical model aimed at simulating the entire population of Sweden.

The studies mentioned above were explicitly labelled as microsimulation, but it is important to note that even calculating simple social statistics such as life expectancy is a method that is similar to microsimulation. In life expectancy calculations, probabilities derived from people who have already died are applied to estimate the likely life-spans of those currently alive. Work for the Joseph Rowntree Foundation [4] (Mitchell *et al*., 2000) incorporated a method similar to microsimulation to ask what would happen were 'Britain more equal' in terms of health. Other work considered how housing market debt might have changed under different scenarios using a simulation based on a million households (Dorling, 1994). Work funded by the ESRC used similarly large data-sets to establish the degree of social polarisation taking place in Britain (Dorling and Woodward, 1996), to examine voting trends (Johnston *et al*., 1998) and the long-term effects of migration (Brimblecombe *et al*., 2000). All of these studies were limited to concentrating on particular issues and did not highlight the fact that the methods being used involved techniques very similar to microsimulation.

It should also be noted that there is currently a major challenge to build on the work described above in order to project the population into the future to predict what would happen under different macro-economic, micro-economic and social policy scenarios. This may enable us, to varying extents, to evaluate the short- and longterm impacts that various government policies are likely to have on different segments of British society and different geographical areas.

Although no serious attempts have been made to simulate *spatially* and *dynamically* the population of Britain, there are numerous examples of successful dynamic microsimulation models being used outside Britain. For instance, Caldwell *et al.* (1998) describe CORSIM, which is a dynamic microsimulation model, under development at Cornell University since 1986 (Caldwell and Keister, 1996; Caldwell *et al.*, 1998). CORSIM has been used to model wealth distribution in the US over the historical period 1960–1995 and to forecast wealth distribution into the future. Another example of dynamic spatial microsimulation is the ongoing work in the Netherlands described by Hooimeijer (1996) which adopts a spatial microsimulation approach to analyse the linkages between supply and demand in the housing market and labour market simultaneously. In the context of this work, the spatial mobility of households is modelled in three different time-sets (daily commuting, relocation, lifetime mobility). The fundamental characteristic of this methodology is the life-course approach to understanding the behaviour of households (Hooimeijer, 1996).

As noted above, a more recent example of comprehensive dynamic spatial microsimulation modelling is the work of the SMC in Sweden (Holm *et al.*, 1996; Vencatasawmy *et al.*, 1999). The SMC built on previous microsimulation modelling efforts (Holm *et al.*, 1996) to construct TOPSIM (Total Population Simulation Models) and *SVERIGE* (System for Visualising Economic and Regional Influences Governing the Environment) which simulate the entire population of Sweden. In addition, *SVERIGE* is aimed at studying the spatial consequences of various national, regional and local-level public policies. The database used for this model comprises longitudinal socio-economic information on every resident of Sweden for the years 1985 to 1995. Furthermore, Ballas *et al*. (2001) argued the need for a spatial microsimulation framework for rural policy

analysis in the Republic of Ireland, and presented a dynamic model of forecasting demographic characteristics at the small area level in the rural localities of the Republic of Ireland. They validated their dynamic model by comparing forecasted figures at the Irish county level with actual data from the 1991 and 1996 Censuses of the Irish Republic's population. Wiemers *et al.* (2003) further extended this modelling work and validated it with the use of the 2002 Irish Republic Census data. It should be noted that the results from this ongoing dynamic microsimulation work have been extremely encouraging (with maximum error margins of 4% for key variables).

Although there are no examples of dynamic spatial microsimulation in the UK, a great deal of dynamic microsimulation work has been conducted at the *national level*. For example, Falkingham and Lessof (1992) presented LIFEMOD, which is an example of a dynamic cohort microsimulation model, simulating the life histories of a cohort of 2000 males and females. Each individual 'experiences' major life events such as schooling, marriage, childbirth, children leaving home and employment. Another example of a UK national dynamic microsimulation model is PENSIM (Hancock *et al.*, 1992) which aims to study the influences of policy change on the income distribution of pensioners up to 2030. More recent work of Falkingham and colleagues involved the use of much larger samples.[5]

All the studies mentioned above suggest that spatial microsimulation is a methodology with huge potential for the modelling and forecasting of small area populations. In the remainder of this paper we report on how we have been building on the experience obtained through these studies in order to create a dynamic spatial microsimulation model for Britain.

THE *SIMBRITAIN* MODEL

Introduction

In this section we present a spatial microsimulation model which aims at simulating the population of Britain into the future. Firstly, we report progress on a project that aims at developing a spatial microsimulation model of the population of the city of York. This model is being used as a pilot study for the application of the technique over the entire country. The data that are being used in the initial stage of the model development are as follows:

- The 1991 UK Census Small Area Statistics (SAS) consisting of a subset of 86 demographic and socio-economic data tables for Great Britain. In particular, the SAS tables contain approximately 9000 statistical counts and are available down to ED (enumeration district) level in England and Wales and output area (OA) in Scotland. Similar data are available for 1981 and 1971.

- The British Household Panel Survey (BHPS), which is a national annual survey of the adult population of the UK, drawn from a representative sample of over 5000 households. The survey collects information on a wide range of variables covering most aspects of life in Britain.

The main advantage of the SAS data is that they provide at the small area level information based on the 1991 UK Census of Population, which was the most authoritative social accounting of people and housing in Britain of its time (Dale and Marsh, 1993). However, the Census records demographic and socio-economic information at a single point in time and, therefore, is less appropriate for the study of social and economic change through time. On the other hand, the BHPS is a very useful tool for the understanding and analysis of social and economic change at the individual and household level. Nevertheless, it has a relatively small sample size and does not take into account the geographical dimension of social and economic household dynamics in Britain. One of the objectives of *SimBritain* is to add a geographical dimension to panel data from surveys such as the BHPS (and to add 'lifehistories' to Census data). In order to do so we have at our disposal a wide range of methodologies. In particular, there are probabilistic synthetic reconstruction methodologies, which aim at using data from different small area statistics tables in order to estimate joint conditional probabilities. Population microdata are then synthesised on the basis of these probabilities. Synthetic reconstruction methodologies are suitable when there is a lack of good quality microdata source. Among the first examples of synthetic reconstruction approaches to the estimation of small area microdata were the work of Duley *et al.* (1988),

Birkin and Clarke (1988, 1989) and Williamson (1992). When small area microdata are available it is possible to reweight them so that they would fit small area statistics tables. Williamson *et al.* (1998) explored different solutions to finding the combination of household Samples of Anonymised Records (SARs) 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*. In addition, Voas and Williamson (2000) further tested and validated these combinatorial optimisation methodologies. Moreover, Ballas *et al.* (1999) and Ballas (2001) also adopted a simulated annealing-based approach to generating small area microdata with the use of the SARs.

It should be noted that the above methodologies involve the use of random sampling procedures. In this paper we present a deterministic approach to reweighting the BHPS households so that they fit given small area statistics tables. A particular characteristic of this method is that it does not use random number generators at any stage (hence the term deterministic), and it therefore produces the same results with each run. Further, we developed a projection methodology that aims at providing estimates of future SAS tables, which, as will be seen, are very close to official government projections of population trends. Moreover, using the reweighting approach we produced BHPS-based population microdata for all wards in Britain.

Data Sources

The main datasets that we used in this project were the British Household Panel Survey (BHPS) and the 1971, 1981 and 1991 UK Censuses of Population. The BHPS is an annual survey of the adult population of the UK, drawn from a representative sample of over 5000 households. The aim of the survey is to deepen the understanding of social and economic change at the individual and household level in Britain, as well as to identify, model and forecast such changes and their causes and consequences in relation to a range of socio-economic variables (Taylor *et al.*, 2001). Tables 1 and 2 outline the core household and individual questions asked in the BHPS questionnaires. These questions have generated a wealth of socio-economic and demographic variables, which make the BHPS unique in that it contains almost all the variables contained in most cross-sectional national social survey data in Britain.

However, one of the drawbacks of the BHPS is that it gives information at relatively coarse levels of geography. In the context of this study we used data from the UK Census of Population in order to add a geographical dimension to the BHPS microdata. The UK Census of Population provides the most authoritative social accounting of people and housing in Britain and is a unique source of data for the social sciences (Dale and Marsh, 1993). Table 3 lists the questions asked by the 1991 UK Census of Population. It should be noted that Census crosstabulations from the SAS provide very useful information on the attributes of the population at the small area level. However, policy-makers

Table 1. Examples of the core question subject areas from the BHPS Household Questionnaire.

Size and Condition of Dwelling	*Household Finances*:
Ownership Status	Rent and Mortgage, Loan and HP Details
Length of Tenure	Local Authority Service Charges
Previous Ownership	Allowances/Rebates
Interview Characteristics	Difficulties with Rent/Mortgage Payments
	Household Composition
	Consumer Durables, Cars, Telephones, Food
	Heating/Fuel Types, Costs, Payment Methods
	Non-monetary poverty indicators
	Crime

Source: Taylor et al. (2001)

Table 2. Details of the core, rotating core and variable component question subject areas from the BHPS Individual Questionnaire.

Core	*Neighbourhood and individual*: Demographics Birthplace, Residence Satisfaction with Home/Neighbourhood Reasons for Moving Ethnicity Educational background and attainments Recent Education/Training Partisan support Changes in marital status Citizenship	*Current Employment*: Employment status Not working/Seeking work Self Employed Sector Private/Public SIC/SOC/ISCO Nature of Business/Duties Workplace/Size of Firm Travelling Time Means of Travel Length of Tenure Hours worked/Overtime Union Membership Prospects/Training/Ambitions Superannuation/Pensions Attitudes to work/Incentives Wages/Salary/Deductions Childcare provisions Job search activity Career Opportunities Bonuses Performance related pay	*Finances*: Incomes from: Benefits/Allowances/Pensions/Rents/Savings/ Interest/Dividends Pension Plans Savings and Investments Material well-being Consumer Confidence Internal Transfers External Transfers Personal Spending Roles of partners/Spouses Domestic work/Childcare/Bills/Everyday Spending Car Ownership/Use/Value of Car Interview Characteristics Windfalls
Rotating Core	*Health and Caring*: Personal health condition Employment constraints Visits to doctor Hospital/Clinic use Use of Health/Welfare Services Social Services Specialists Check-ups/Tests/Screening Smoking Caring for relatives/others Time spent caring for others Private medical insurance Activities in daily living	*Employment History*: Past year Labour Force Status Spells Size/Sector/Nature of Business/Duties Wages/Salary/Deductions Reasons for leaving/taking jobs	*Values and Opinions*: Partisanship/Interest in Politics Religious Involvement Parental Questionnaire
Variable Components	*Lifetime Marital Status History (Wave 2)*: Number of marriages Marriage dates Divorce/widowhood/ Separation dates Cohabitation before marriage *Lifetime Marital Status History (Wave 3)*: Start and finish dates Labour force status Sector/nature of business duties *Health and Caring*: Children's health Other Health Scales: SF36 (Wave 9) *Computers and Computing (Wave 6/7)*: Ownership and usage	*Lifetime Fertility and Adoption History (Wave 2 and Wave 8 catch-up)*: Birth dates Adoption dates Sex of children Leaving or mortality dates *Lifetime Cohabitation History (Wave 2 and Wave 8 catch-up)*: Start and finish dates Number of partners *Neighbourhood and Demographics*: Driving Licence Parents employment background Family background Difficulties with debt Community and Neighbourhood *Employment (Wave 9)*: National Minimum Wage Work strain Work orientation	*Lifetime Employment Status History(Wave 2)*: Start and finish dates Employment status *Values and Opinions*: Aspirations for children Important Events Quality of Life *Credit and Debt*: Investment and Savings Commitments *Crime*: Criminal activity on local area Perceptions of crime

Source: Taylor et al. (2001)

Table 3. 1991 UK Census questions (after Openshaw, 1995: 2).

Principal 1991 Census questions
Household questions
Type of accommodation
Number of rooms
Tenure
Amenities
Car and van ownership
Individual person questions
Name
Sex
Date of birth
Marital status
Relationship in household
Whereabouts on census night
Usual address
Term-time address
Usual address one year ago
Country of birth
Ethnic group
Long-term illness
Whether working, retired, looking after house etc.
Hours worked per week
Occupation
Name and address of employer
Address of place of work
Daily journey to work
Degree, professional and vocational qualification

are often interested in cross-tabulations of variables that are not available from the SAS. This problem can be overcome if Census microdata are available. It should also be noted that population individual-level data or microdata are a valuable resource for social science research. Compared with aggregate tabular population data, such as the SAS, population microdata contain much more detail on household or individual attributes, but are released at a coarser geographical level. The growing need for population microdata has led an increasing number of governments to commit to the decennial production of census-based microdata samples. In Britain, the release of microdata files had been under discussion since the 1970s, and after a detailed assessment of the likely risk to confidentiality, the UK Census offices agreed to release Samples of Anonymised Records (SARs) from the 1991 Census of Population (Marsh and Teague, 1992; Marsh, 1993; Middleton, 1995). Sadly, 2001 SARs are not currently planned to be released in similar detail. One of the major limitations of Census microdata is that they are constrained by the number of questions asked in the Census. It can be argued that the creation of a data-set that had the geographical detail of Census data and the wealth of information that is contained in the BHPS would be extremely useful and policy-relevant. In the next sections we describe the methodology that we developed and used to create such a data-set for the city of York and subsequently for the whole of Britain for the years of 1991, 2001, 2011 and 2021.

Projecting Small Area Statistics into the Future

In order to project the population of Britain into 2001, 2011 and 2021 we used data from previous Censuses. In particular, projections of a set of small area statistics tables (described in Table 4)

Table 4. Constraint tables.

Table	Category		
	1	2	3
Car Ownership	No cars	1 car	2+ cars
Class Composition	Affluent	Middle-class	Less affluent
Demography	1 child	2+ children	No children
Employment	Economically active	Retired	Inactive
Household Composition	Married couple	Lone parent	Other
Tenure	Owner occupied	Council tenants	Other

were calculated using the 1971, 1981 and 1991 Census Small Area Statistics (SAS). Using these three time points, a trend curve was produced allowing tables to be predicted up to 2021. In particular, we first considered a 'gravity' model for projecting constraint variables of the form:

$$A = \exp(\ln W * (\ln w)^2 * \ln u / (\ln v)^3 \quad (1)$$

where *u*, *v* and *w* are the smoothed proportions in 1971, 1981 and 1991 respectively, *W* is the observed ward proportion in 1991 and *A* is the projected ward proportion in 2001. Nevertheless, concerns that this model was unduly affected by inflexions in the data – for instance, a decline in the value of a proportion during the period 1971–81 followed by a recovery during 1981–91 led to unrealistically high estimates for 2001 – subsequently led us to investigate various other possible models for use with a time series of proportions. One methodology that seems to produce more conservative projections while still fitting the 1971–91 data is offered by Holt's linear exponential smoothing (Holt, 1957). This is an extension of exponential smoothing to take into account a possible linear trend. There are two smoothing constants α and β. The equations are:

$$L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + b_{t-1})$$
$$b_t = \beta (L_t - L_{t-1}) + (1 - \beta) b_{t-1}$$
$$F_{t+m} = L_t + b_t m$$

where L_t and b_t are respectively (exponentially smoothed) estimates of the level and linear trend of the series at time *t*, whilst F_{t+m} is the linear forecast from *t* onwards. Some initial work has been undertaken to test the relative performance of the gravity and exponential smoothing models. The findings suggest that exponential smoothing avoids some of the less credible forecasts from the gravity model. This is partly because the model is intrinsically more conservative, but also reflects the fact that the choice of values for α and β is under user control. On that basis we undertook projections of small area statistics tables at the ward level. In order to avoid problems of wards with low populations, the ward data were smoothed. For the tables for each ward we took the change that has occurred for the 20,000 population nearest to that ward. This change was then applied to the values for each ward.

Each table has three categories of households. Projections were calculated using proportions of households in each category in each ward. The first category was calculated as in the equations below; the second category was then calculated as the proportion that category takes up of what is left, using the following equation:

$$W_{t+10} = \exp[\ln W_t (\ln w_t)^2 (\ln w_{t-20} / (\ln w_{t-10})^3)] \quad (2)$$

where *W* is the ward proportion, *w* is the smoothed ward proportion and *t* is the census year. The proportion in the third category was then calculated as 1 minus the proportions in the two other categories, to ensure that all households were allocated to a category.

Once projections of the proportions of households that fall into each table category were calculated they needed to be applied to ward-level household projections to produce estimates of the numbers of households in each ward in each category. There are no official household projections at ward level; therefore a method had to be devised. Projections were required every 10 years between 1991 and 2021. The Office for the Deputy Prime Minister (ODPM) publishes projections of total households for counties for every five years up to 2021. These county totals

were used to produce ward-level estimates of households. It was assumed that the distribution of households between wards in each county would remain the same as in 1991. This 1991 distribution can be calculated from the 1991 Census of Population. For each county the proportion of its households in each ward is calculated. This proportion is assumed to remain stable up to 2021. Therefore ward-level household projections can be produced by multiplying the ODPM household projections for counties by the proportion of households for each ward.

Six constraint tables were created, each with three categories. The tables and their categories are listed in Table 4.[6] The variables for these tables were selected on the basis of continuity throughout the three Censuses that we used in the context of this project (i.e. we had to use variables that were measured in all three Censuses).

Reweighting the BHPS

Having estimated small area statistics tables for 2001, 2011 and 2021, the next task was to generate small area population microdata for these years. In order to do so we used existing microdata from the BHPS. In particular, we calculated the appropriate weights for all BHPS households for each simulated geographical area, so that they would fit the Small Area Statistical descriptions described above. It should be noted that all BHPS households have been given a weight that compensates for error, bias, refusals, and so on. In particular, in the BHPS, household weights were applied to compensate for the unequal selection probability arising from the two-stage stratified sampling design, to compensate for nonresponding households, and to adjust for those individuals in a responding household who failed to give a full interview (Taylor ***et al***., 2001). One of the tasks that we faced in this project was to readjust the original weights of BHPS households so that the new weights would add up to small area constraints such as those described in the previous section.

As noted above, we adopted a deterministic reweighting approach to readjust the given BHPS household weights so that when all household weights are added up they fit the small area constraints described above. This is described in Tables 5–8. In particular, Table 5 gives a hypothetical individual microdata-set comprising five individuals, which fall within two age categories. Further, Table 6 depicts a small area statistics table for a hypothetical area, whereas Table 7 depicts a cross-tabulation of the hypothetical microdata-set, so that it can be comparable to Table 6.

Table 5. A hypothetical microdata set.

Individual	**Sex**	**Age-group**	**Weight**
1st	Male	Over-50	1
2nd	Male	Over-50	1
3rd	Male	Under-50	1
4th	Female	Over-50	1
5th	Female	Under-50	1

Table 6. Hypothetical small area data tabulation.

Age/sex	Male	Female
Under-50	3	5
Over-50	3	1

Table 7. The hypothetical microdata-set, crosstabulated by age and sex.

Age/sex	Male	Female
Under-50	1	1
Over-50	2	1

Using these data it is possible to readjust the weights of the hypothetical individuals, so that their sum would add up to the totals given in Table 6. In particular, the weights can be readjusted by multiplying them by the value in the cell in Table 7, which denotes the category to which they belong, over the respective cell in Table 8. This can be expressed as follows:

$n_i = w_i * s_{ij}/m_{ij}$ (3)

where n_i is the new household weight for household i, w_i is the original weight for household i, s_{ij} is element ij of table s (small area statistics table, which is the equivalent of table 6) and m_{ij} is element ij of table m (reproduced table using the household microdata original weights, which is

the equivalent of Table 7 in our example). Table 8 depicts how this simple formula is used to readjust the weights of the individuals in our example. The above process can then be used to reweight the individuals to fit another table.

Table 8. Reweighting the hypothetical microdata-set in order to fit Table 6.

Individual	Sex	age-group	Weight	New weight
1st	Male	Over-50	1	1 * 3/2 = 1.5
2nd	Male	Over-50	1	1 * 3/2 = 1.5
3rd	Male	Under-50	1	1 * 3/1 = 3
4th	Female	Over-50	1	1 * 1/1 = 1
5th	Female	Under-50	1	1* 5/1 = 5

Table 9. Origin of wave 1 BHPS households (AREGION).

Value Label	Frequency	Frequency (%)
Inner London 1	498	5.8
Outer London 2	597	7
Rest of South East 3	1611	18.9
South West 4	713	8.4
East Anglia 5	303	3.6
East Midlands 6	595	7
West Midlands Conurb 7	391	4.6
Rest of West Midlands 8	369	4.3
Greater Manchester 9	396	4.6
Merseyside 10	195	2.3
Rest of North West 11	363	4.3
South Yorkshire 12	197	2.3
West Yorkshire 13	299	3.5
Rest of Yorks & Humber 14	257	3
Tyne & Wear 15	202	2.4
Rest of North 16	293	3.4
Wales 17	392	4.6
Scotland 18	853	10

In the context of this project we adopted this reweighting procedure iteratively to readjust the BHPS household weights so that they would fit the small area statistics tables described in the previous section. The generated weights for each household represent the probabilities of BHPS households to 'reside' in a given small area.

One of the difficulties encountered with the reweighting methodology described above was the high presence of BHPS households coming from geographical areas other than the simulated area (in particular, there was a high presence of households from the southeast of England). Table 9 shows the geographical distribution of the households in the BHPS wave 1. As can be seen, around 33% of the households come from the South East, whereas only about 10% of households come from Yorkshire and Humberside.

In the case of the simulation of the population in York, the initial geographical distribution of the BHPS households would result in the selection of large numbers of non-Northern households from wave 1 that would populate the York wards. In order to deal with this problem we explored a number of possible solutions. One of the solutions explored was the development of 'geographical multipliers' that could be used to increase the chances of 'local' BHPS households being included in the simulation. Under this approach all BHPS records were used to create the matrix shown in Table 10. In order to do so, each record (individual) in the BHPS database and the category with regards to Table 10 was identified. Then the respective table cell was incremented by the quotient of 1 over the number of household members (e.g. for a 4-member household, the cell would be incremented by 0.25).

Table 10. Geographical multipliers.

Place of birth/ place of residence	Inner London 1	Outer London 2	R. of South East 3	South West 4	…	R. of Yorks & Humber 14	…	Scotland 18
C/Ldn.C/Wminster	0.36	0.12	0.34	0.08	…	0.00	…	0.01
Camden	0.21	0.27	0.25	0.12	…	0.00	…	0.00
Hackney	0.13	0.32	0.32	0.10	…	0.00	…	0.00
Hammsmith/Fulham	0.31	0.20	0.26	0.10	…	0.02	…	0.00
Haringey	0.12	0.30	0.36	0.16	…	0.00	…	0.00
Islington	0.22	0.25	0.31	0.10	…	**0.02**	**…**	0.00
Knsingtn/Chelsea	0.14	0.09	0.28	0.36	…	0.00	…	0.00
Lambeth	0.29	0.29	0.22	0.09	…	0.01	…	0.01
Lewisham	0.34	0.17	0.28	0.06	…	0.00	…	0.00
.	.	.	.	.	…	.	…	.
.	.	.	.	.	…	.	…	.
.	.	.	.	.	…	.	…	.
Glasgow City	0.04	0.01	0.10	0.01	…	**0.02**	**…**	**0.70**
Motherwell	0.00	0.00	0.00	0.04	…	0.04	…	0.87
Renfrew	0.04	0.00	0.00	0.00	…	0.00	…	0.52
Angus; Perth & Kinross	0.00	0.09	0.03	0.00	…	0.00	…	0.82
Dundee City	0.00	0.00	0.01	0.00	…	0.00	…	0.96

Once the table above is filled in, the new weight multipliers are calculated by dividing the total number of people residing in a given region by the total number of people that have the same birthplace.

After having calculated all the weight multipliers, the next step is to use them in order to change the BHPS weights. For simulating York,[7] only the column 'Rest of Yorks and Humberside 14' of the above table is needed. All BHPS households are read and they are assigned a multiplier based on the birthplace of the head of household. For example, if the head of household was born in Islington in the above table, the household would get a multiplier of 0.02; similarly a household with a head born in Glasgow would also be assigned a multiplier of 0.02 (see Table 10). However, if we were simulating a Scottish city, then the multipliers from the last column (Scotland 18) would be used and therefore a household with a head born in Glasgow would have a much larger multiplier of 0.70. However, one of the problems associated with this methodology is the effect that the increase of chances of 'local' households being included in the simulation were cancelled out when the reweighting procedure described above was implemented.

An alternative geographical weighting methodology that we developed and tested involved the creation of a 'geographical' constraint that would increase the chances of 'local' BHPS households being selected for simulation. The creation of this constraint was based on the birthplace of BHPS heads of household. Table 11 shows the data used to create this constraint. In particular, Table 11 is a cross-tabulation of the residence by birthplace of BHPS heads of households (both at the geographical level of the BHPS region).

The data shown in Table 11 were used to generate a table of geographical multipliers for each region. These geographical multipliers represented the heads of households born in a given region as a proportion of the total household heads living in the region. For instance, Table 12 shows the geographical multipliers for the 'Rest of Yorkshire and Humberside' region.

The data shown in Table 12 were used to create a geographical constraint table for each small area in the 'Rest of Yorkshire and Humberside'. Table 13 shows the geographical constraint table for the Clifton ward of York. This table was created by multiplying the total number of households counted in Clifton in 1991 by the respective birthplace multipliers shown in Table 12.

Table 12. Geographical multipliers table for Rest of Yorkshire and Humberside.

Birthplace of household heads	Count	Multiplier
Inner London	1	0.00585
Outer London	0	0.00000
South East (rest)	5	0.02924
South West	4	0.02339
East Anglia	2	0.01170
East Midlands	6	0.03509
West Midlands MA	4	0.02339
West Midlands (rest)	0	0.00000
Greater Manchester MA	4	0.02339
Merseyside MA	2	0.01170
North West (rest)	7	0.04094
South Yorkshire MA	12	0.07018
West Yorkshire MA	9	0.05263
Yorkshire & Humberside (rest)	86	0.50292
Tyne & Wear MA	2	0.01170
North (rest)	8	0.04678
Wales	2	0.01170
Scotland	6	0.03509
Northern Ireland	2	0.01170
Abroad	5	0.02924
Missing	4	0.02339
Total	**171**	**1.00000**

Table 13. Geographical constraint table for Clifton, York.

Birthplace	Clifton
Inner London	16.04
Outer London	0.00
South East (rest)	80.20
South West	64.16
East Anglia	32.08
East Midlands	96.25
West Midlands MA	64.16
West Midlands (rest)	0.00
Greater Manchester MA	64.16
Merseyside MA	32.08
North West (rest)	112.29
South Yorkshire MA	192.49
West Yorkshire MA	144.37
Yorkshire & Humberside (rest)	1379.52
Tyne & Wear MA	32.08
North (rest)	128.33
Wales	32.08
Scotland	96.25
Northern Ireland	32.08
Abroad	80.20
Missing	64.16
Total	**2743**

As can be seen, Table 13 sets the constraint that 1379.52 of the BHPS households selected to populate Clifton should have a head of household born in the 'Rest of Yorkshire and Humberside' (i.e. Yorkshire and Humberside excluding West and South Yorkshire). The geographical constraint tables were added to the small area statistics tables described above, before implementing the reweighting procedure also described above. Thus all the BHPS households were reweighted so that they would fit the following tables:

- Table 1: Birthplace (18 categories)
- Table 2: Car ownership (3 categories)
- Table 3: Demography (3 categories)
- Table 4: Household type (3 categories)
- Table 5: Tenure (3 categories)
- Table 6: Economic activity and socio-economic group (5 categories)

However, this methodology led to large overestimates or underestimates in certain areas of variables that were not used in the constraining procedure. Having experimented with these methodologies we concluded that the best approach was to define the BHPS sample used in the simulation on the basis of the geographical area being simulated. For instance, in the simulation of York we used only the BHPS households that lived in the BHPS region 'Rest of Yorkshire and Humberside' (AREGION = 14). After generating the BHPS household weights for each ward in York, the next step was to select the appropriate households (or, in other words, convert the decimal weights or probabilities into integer weights). In the context of this project we developed and tested different 'integer weighting' or *integerisation* methodologies. The first method we developed and tested involved giving all households an integer weight equal to rounding of their decimal weight. For instance, a household with a weight of 2.4 would be given a household weight of 2. This led to an initial selection of households, which was, however, less than the total number of households in a given geographical area. In order to select the remaining household the following algorithm was implemented:

(1) Set a *rounding* variable equal to 0.999.
(2) Increase by 1 the weight of all households that have a decimal remainder bigger than or equal to *rounding*.
(3) Decrease *rounding* by 0.0001.
(4) If sum of all weights is equal or bigger than total number of households in the area, then exit.
(5) Return to step 1.

It should be noted, however, that the above methodology minimises the probabilities of households that have a decimal remainder close to 0 of being selected. In order to ameliorate this problem alternative strategies can be adopted. For instance, an alternative method that we tested involved assigning integer weights on the basis of random sampling. In particular, it is possible to treat the household weights as a distribution of probabilities and to perform Monte Carlo sampling from this distribution to select a number of households that would be equal to the given small area total. Nevertheless, as noted above, the overall modelling approach that we adopted was deterministic and we therefore wanted to avoid the use of random number generation procedures. We therefore devised an alternative deterministic approach to integer weighting which was based on the following methodology:

Define two variables named ***counter*** and ***weight*** and set them to zero, and then:

(1) Sort all households into ascending order of probability of living in the small area (which were calculated using the method described above) being populated.
(2) Increase cumulative ***weight*** by the weight (probability) of the next sorted household **h(*counter*)**. For instance, if ***counter*** = 0, the ***weight*** is increased by the probability of the first household: h(0).
(3) If cumulative ***weight*** > 1, give to the household ***h(counter)*** an integer weight equal to the rounded *weight* value and subtract this value from ***weight*** (e.g. if ***weight*** = 2.05 set household ***weight*** = 2 and set ***weight*** = 2.05 - 2 = 0.05). Increase ***counter*** by 1 (move to next household).
(4) If ***counter*** < total number of households in the small area, return to 2, else exit.

The implementation of the above algorithm led to the creation of a ward-level microdata-set for the city of York.[8] Clearly, the integerisation process described above resulted in an increase of the difference between the 'simulated' and actual cells of the target variables described in Table 4 (the following section presents in more detail these differences for a selection of variables). Furthermore, we observed that, when comparing weighted data with their 1991 Census counterparts, there were, in some wards, relatively high overestimates and underestimates of some variables that were not used as constraints in the simulation. In order to tackle this problem we developed an algorithm aimed at swapping suitable simulated households between wards in order to reduce the error further. The steps taken were as follows:

(1) Identify wards with the highest overestimates and underestimates for each variable.
(2) Compare each household in the simulated database with all other households and search for households that have all attributes (used as constraints, listed in Table 4) in common but one.
(3) For each pair of almost identical households, swap the households between the areas with the highest overestimate and underestimate.
(4) Move to the next household and repeat the process.

The implementation of this algorithm led to the reduction of error for unconstrained variables in some areas. One difficulty with the use of this algorithm was that there were no available census data for the years 2001, 2011 and 2021. In order to tackle this problem we applied the variable rates from 1991 to the estimated number of individuals for these years. It should be noted that the analysis could be improved by projecting these rates into the future on the basis of past trends. The following section shows how *SimBritain* can be validated through the comparison of simulated data with actual population statistics.

Validation

One way of checking the reliability of our projection methodology is by using past Census data to project distributions of populations into 1991, and then compare the projected values with the actual data from the 1991 Census. Table 14 shows an example of a comparison of Census data

Table 14. Comparing Census data with projected data for 1991 in York (projection based on data from the Censuses of 1961, 1971 and 1981).

	Census data						
Year	**1951**	**1961**	**1971**	**1981**	**1991**	**Predicted proportion for 1991**	**Difference between projection and actual data**
Class I & II	19%	21%	24%	28%	34%	34%	0%
Class III	51%	50%	49%	47%	43%	44%	1%
ClassIV & V	30%	29%	27%	25%	24%	22%	-2%

on social class groupings and projected proportions of these groups in 1991. As can be seen, by using the data on social class for the years 1961–71–81, our projection method predicts that 34% of the households in York in 1991 would belong to classes I and II. This prediction matches the actual proportion calculated from 1991 Census data. Likewise, our projection method works well in estimating the 1991 distributions of class III and class IV & V households. Nevertheless, there is a certain degree of variation in the performance of this projection method. Table 15 compares Census data on social classes IV and V with projected proportions of these groups in 1991 for all local authorities in Wales. As can be seen, the projection method performs relatively well for most areas, with the exception of a few local authorities. In particular, it is noteworthy that in more than half of the local authorities in Wales the absolute differerence between actual and projected ten-year forward proportions are equal to or less than 5%.

The next step in the validation procedure is to test the results from the reweighting exercise described above. As noted above, we used a reweighting methodology to select the BHPS household that best matched six small area Census variables. The 1991 Census provides an obvious benchmark against which to test the results from *SimYork* and *SimBritain*.[9] Firstly the composition of each of the 641 British parliamentary constituencies is simulated using the methodology outlined previously. Secondly, a set of variables is identified which is present both in the BHPS and the Census. Thirdly, the constituency- level average of each of these variables is determined from our simulated population and from the actual Census counts. Finally these values are compared to see how closely *SimBritain* matches *RealBritain*.

A total of eight variables was chosen for the comparison. These are listed in Table 16 together with their average values in the BHPS and in the Census. Several points are worth noting. Firstly, as most Census variables are counts, the majority of comparisons relate to proportions. The one exception, average age, assumes an even distribution of ages within each of the Census age categories, but this is not unreasonable. Secondly, the average values of six of the eight variables match very closely and the exceptions do not differ to such a degree as to make comparisons invalid. Nevertheless, it would be worth investigating why the average values for unemployment and migration differ by 2.0% and 1.8% respectively. Thirdly, the variables selected combine various aspects of household and individual characteristics. Age is an individual-level characteristic available for all residents including children. Illness, unemployment, education and travel-to-work are all measured at the individual level but relate to different sections of the adult population. Central heating is measured at household level in both the Census and BHPS, whereas values for migration and ethnicity for *SimBritain* have to be calculated from a combination of individual and household records.

The degree to which the values from *SimBritain* and *RealBritain* match on these variables will depend upon a number of factors. Firstly, do they measure the same thing? As noted above, the match is generally good, but there are some concerns regarding unemployment and migration. Secondly, how closely related is each variable to the constraint variables? As households have been selected for each constituency on the basis of their values on the constraint variables, a variable which is closely related to those constraints stands a better chance of being well

Table 15. Comparing Census data with projected data for 1991 in Welsh Local Authorities (projection based on data from the Censuses of 1961, 1971 and 1981).

Social classes IV and V						61-71-81	difference
Year	1951	1961	1971	1981	1991	projected 1991	
Aberavon	37%	33%	29%	32%	29%	40%	11%
Alyn and Deeside	36%	33%	30%	26%	24%	21%	-3%
Blaenau Gwent	41%	35%	30%	31%	30%	36%	6%
Brecon and Radnorshire	34%	30%	27%	25%	24%	25%	1%
Bridgend	31%	27%	23%	23%	22%	27%	4%
Caernarfon	34%	30%	26%	25%	24%	26%	2%
Caerphilly	36%	29%	24%	27%	24%	39%	15%
Cardiff Central	28%	25%	23%	22%	20%	24%	4%
Cardiff North	27%	24%	21%	20%	18%	21%	3%
Cardiff South and Penarth	29%	25%	22%	22%	19%	23%	4%
Cardiff West	29%	26%	23%	23%	21%	26%	5%
Carmarthen East and Dinefwr	32%	28%	25%	23%	22%	22%	0%
Carmarthen West and S Pembs	32%	29%	27%	22%	21%	17%	-4%
Ceredigion	30%	27%	24%	21%	21%	18%	-3%
Clwyd South	36%	32%	29%	25%	24%	21%	-3%
Clwyd West	31%	29%	26%	23%	23%	18%	-5%
Conwy	32%	28%	26%	24%	25%	25%	0%
Cynon Valley	38%	32%	27%	29%	28%	38%	10%
Delyn	37%	33%	30%	26%	23%	21%	-2%
Gower	35%	30%	26%	23%	23%	21%	-2%
Islwyn	41%	34%	29%	30%	28%	39%	11%
Llanelli	35%	31%	27%	24%	24%	22%	-1%
Merionnydd nant Conwy	31%	28%	26%	24%	25%	21%	-4%
Merthyr Tydfil and Rhymney	40%	34%	29%	31%	29%	42%	13%
Monmouth	35%	31%	27%	24%	23%	23%	0%
Montgomeryshire	33%	29%	27%	24%	22%	22%	0%
Neath	36%	32%	28%	29%	26%	34%	8%
Newport East	34%	30%	26%	24%	22%	23%	0%
Newport West	34%	30%	26%	25%	22%	28%	5%
Ogmore	35%	29%	24%	27%	26%	37%	10%
Pontypridd	36%	30%	24%	25%	25%	33%	8%
Preseli Pembrokeshire	33%	30%	27%	23%	21%	19%	-3%
Rhondda	38%	33%	29%	31%	28%	40%	12%
Swansea East	35%	31%	27%	27%	25%	30%	5%
Swansea West	34%	30%	26%	25%	24%	25%	1%
Torfaen	40%	35%	30%	30%	27%	34%	7%
Vale of Clwyd	30%	28%	26%	24%	22%	20%	-2%
Vale of Glamorgan	29%	26%	23%	21%	21%	21%	0%
Wrexham	36%	33%	30%	27%	25%	23%	-2%
Ynys Mon	34%	30%	26%	25%	24%	25%	0%

modelled than one which is remotely related. Thirdly, to what extent do regional or local geographical factors influence each variable? Figures 1 to 4 show the scatterplots for some of the variables described in Table 16, with the Census proportion on the vertical axis and the simulated proportion on the horizontal axis. A perfect match would find all 641 points on a straight line of gradient 1. Three statistics help to measure the degree of departure from that pattern. Firstly, there is the standard error of the estimate, the standard error about the regression line. Secondly, there is R-squared, the proportion of the variance in the Census measure predicted by *SimBritain*.

Table 16. Comparing eight variables between the 1991 Census and the BHPS (wave 1) at the national level.

Variable	BHPS	Census
Average age of residents	37.6	37.5
Proportion of adults with long-term illness	.146	.149
Proportion of wholly moving households	.084	.066
Proportion of economically active unemployed	.072	.092
Proportion of 18+ with higher educational qualifications	.122	.134
Proportion of employed travelling to work by public transport	.160	.157
Proportion of households without central heating	.182	.187
Proportion of households with a non-white head	.038	.040

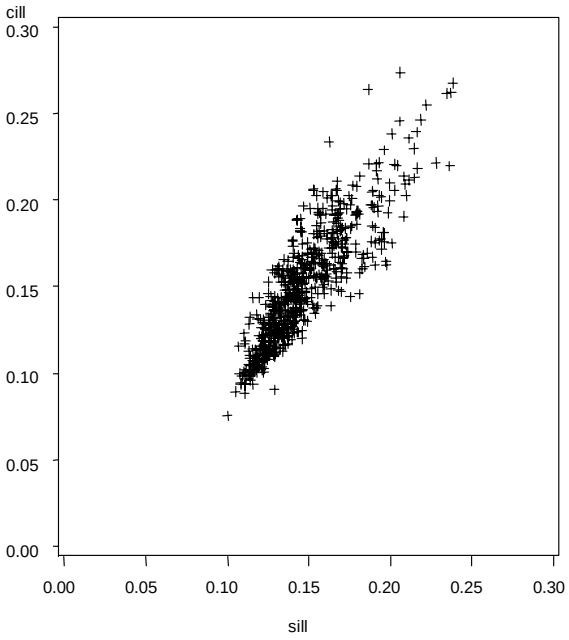

Figure 2. Long-term illness (s.e. = 1.7; R2 = 0.767; beta = 1.19).

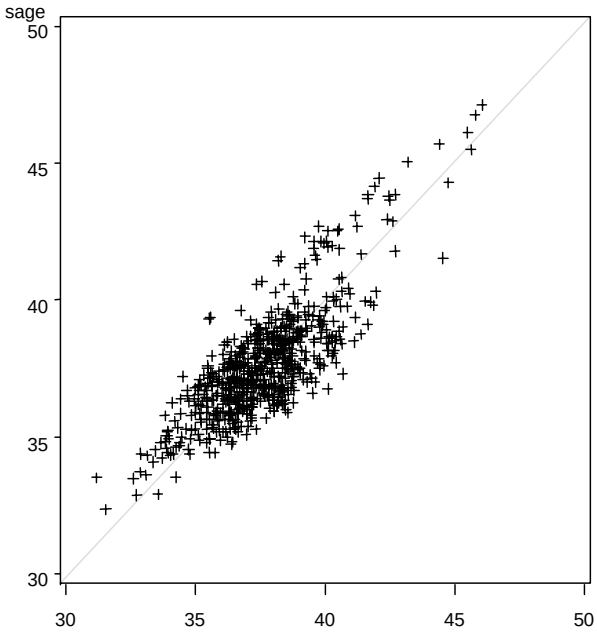

Figure 1. Average age (s.e. = 1.0; R2 = 0.760; beta =1.22). c = census data; s = simulated data (e.g. sage = simulated age; cage = census age).

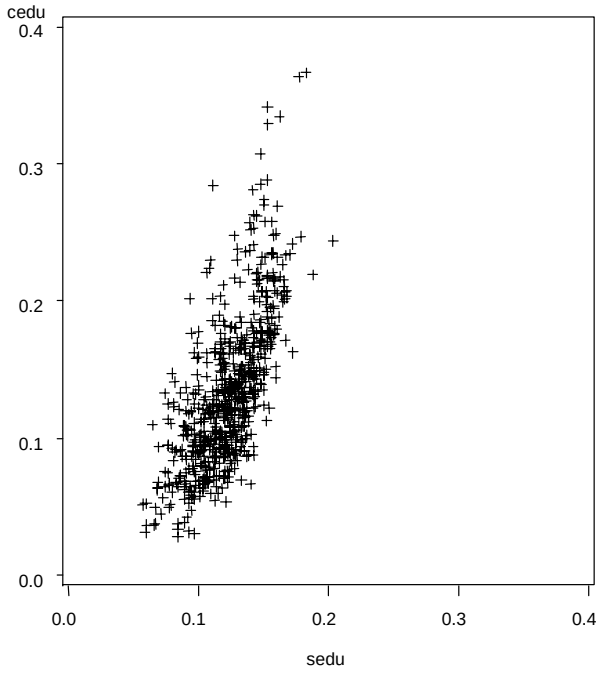

Figure 3. Educational qualifications (s.e. = 4.0; R2 = 0.478; beta = 1.60).

And thirdly there is beta, the slope of the best-fit line through the 641 points. What do the four scatterplots tell us? Dealing with the first two, average age and long-term illness, these demonstrate a reassuring correspondence between *SimBritain* and *RealBritain*. Over three-quarters of the real variation is predicted, and slopes in both cases are reasonably close to unity. However, the model performs less well in the prediction of variables such as educational

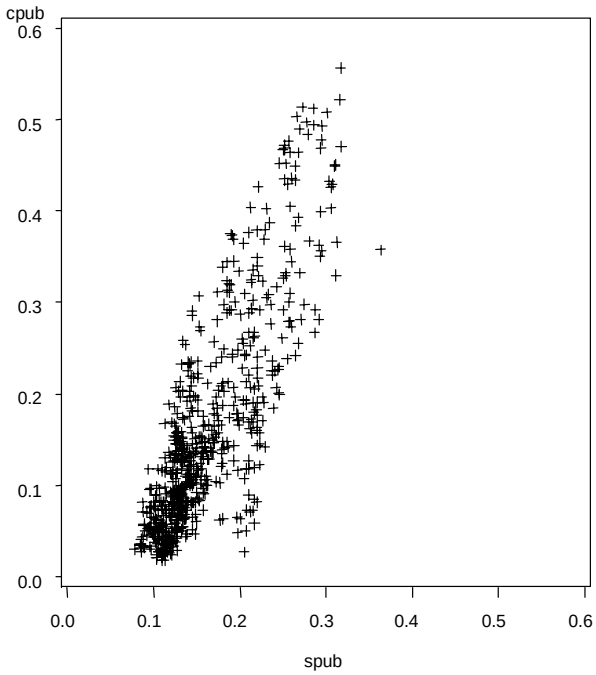

Figure 4. Work by public transport (s.e. = 6.5; R2 = 0.682; beta = 1.76).

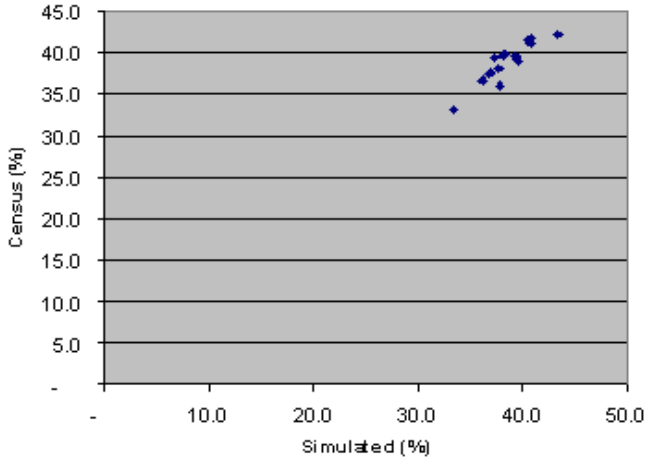

Figure 5. Simulated vs. actual average age of residents by ward, York 1991.

Table 17. Simulated vs. actual Census values for selected variables in York.

Variable	*SimYork* 1991	York (Census 1991)
Average age of residents	38.2	38.6
Proportion of adults with long-term illness	12.1%	12.8%
Proportion of economically active unemployed	4.6%	7.6%
Proportion of 18+ with higher educational qualifications	6.9%	8.5%
Proportion of employed travelling to work by public transport	4.8%	8.0%

qualifications and work by public transport. It was to be expected that the performance of the model would vary from variable to variable. In the case of age and illness the estimates were encouraging, but the other variables were much less predictable.

It is interesting to examine the performance of the model at the small area level. As noted in the introduction, part of *SimBritain* is the *SimYork* model, which aims at exploring simulation results to a greater degree for the city of York. *SimYork* is used as a pilot model, which enables us to evaluate the performance of our simulation for different variables and at various geographical scales.

Table 17 shows the actual (Census) and simulated values for a selection of variables for York. As can be seen, the model underestimates the York unemployment rate, the percentage of the working population who travel to work by public transport, and the percentage of households without central heating. This can be partly explained by the general difference between the Census national rates and the BHPS rates for these variables, and may be due to slight differences in the definitions of these variables.

Figures 5 to 8 show the scatterplot for some of these variables at the ward level, the Census proportion on the vertical and the simulated proportion on the horizontal axis. As can be seen in Fig. 5, there is a relatively good match of simulated and actual values of average age of residents across the 15 wards of York. Nevertheless, as Fig. 6 demonstrates, there is a relatively worse match for the values of actual and simulated values of travel to work by public transport. Likewise, there is a relatively bad match of simulated and actual values of the percentage of adults with higher education, as can be seen in Fig. 7. Finally, there is slightly better match of the simulated and actual ward values of the percentage of individuals reporting limiting long-term illness (Fig. 8).

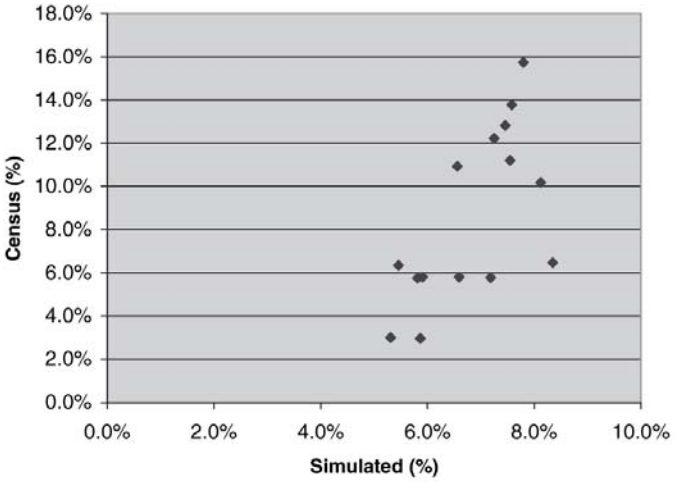

Figure 6. Simulated vs. actual rate of working population travelling to work by public transport (by ward), York 1991.

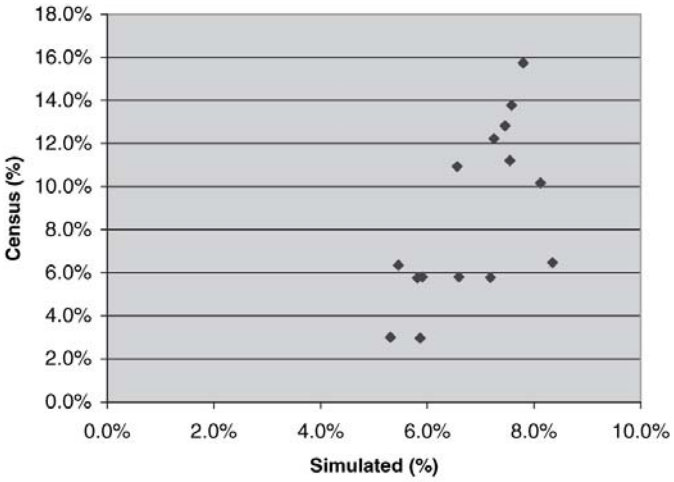

Figure 7. Simulated vs. actual degree or higher degree rate of residents by ward, York 1991.

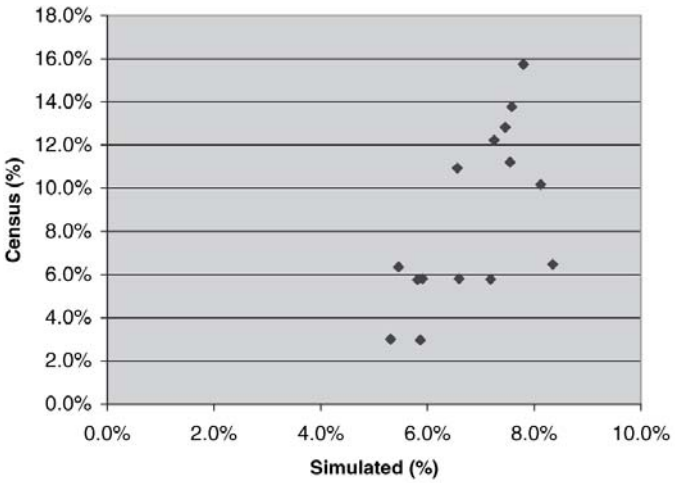

Figure 8. Simulated vs. actual rates of limiting longterm illness by ward, York 1991.

It can be argued that the general underestimation of the percentage of working population who travel to work by public transport may be due to the fact that we are simulating urban localities, where the likelihood of public transport use is higher. Additionally, the large differences between actual and simulated rates of individuals with higher educational qualifications may well be due to the fact that York is a university town.

It would be reasonable to expect that the performance of the model would vary from variable to variable, especially at areas as small as wards and for variables which were not included as constraints in the simulation exercise. Nevertheless, it is encouraging that, as seen in Table 16, the estimates for the city of York match reasonably well their actual Census counterparts. It can therefore be argued that *SimBritain* is relatively reliable when analysing socio-economic patterns at the level of the city. Nevertheless, at the ward level the performance of the model varies considerably, and there is a need to introduce further constraints or data-sets in order to perform analysis at the ward or sub-ward level for particular variables.

CONCLUSIONS

In this paper we reported progress on the *SimBritain* spatial microsimulation model. Firstly, we briefly reviewed past attempts to build spatial microsimulation models and we pointed out that despite the increasing number of such model developments, there have been very few attempts to build dynamic spatial microsimulation models. We then presented the *SimBritain* model, which aims at simulating dynamically the population of Britain at different geographical scales up to the year 2021. We presented methodologies that aim to project small area statistics tables and we

presented a technique that we developed and used to reweight existing survey data so that they would fit small area population descriptions. Finally, we validated our methodologies by comparing simulated with actual data. According to our findings so far, the modelling methodology presented in this paper performs reasonably well at the parliamentary constituency level. Nevertheless, its performance is less satisfactory for certain variables at smaller geographical area levels, such as the electoral ward.

Amongst our immediate priorities is to explore ways of improving the performance of *SimBritain*, which will then be used to estimate and analyse future socio-economic trends in the country. Also, we aim to use the returns of the 2001 Census of Population to thoroughly validate the model across the whole country.

SimBritain will also be used to verify the policy relevance of dynamic spatial microsimulation by simulating alternative *what-would-have-happenedif* scenarios for the period 1991–2001 and to project the population microdata-set in different ways for each year up to 2021 under these futureoriented 'what-if' policy scenarios. Furthermore, *SimBritain* aims to provide an illustration of how a simulation of the whole population of Britain would appear and also gives an indication of the strengths and weaknesses of the dynamic spatial microsimulation approach.

A further goal of our research is to render *SimBritain* capable of simulating the changing population of the whole of Britain into the future, first under the assumption that it continues to change as it has, and then under different scenarios. In particular, alternative projections may also be provided on the basis of proposed government policies and of *hypothetical progressive social policy schemes* that could potentially bring about a major redistribution of income and wealth in British society. One timely issue that we could initially focus on is the eradication of child poverty (for an example of how microsimulation can be used in this area, see Sutherland and Piachaud, 2001). For instance, we aim at rendering *SimBritain* capable of estimating the degree of child poverty eradication within the next 20 years under different policies and assumptions, in order to suggest where policies may be failing in their goal of eradicating child poverty within a generation. In addition, *SimBritain* could also be used to assess the impact of different social policy options on the future demand for pensions, health and personal social services, and longterm care.

ACKNOWLEDGEMENTS The work reported in this paper is funded by the Joseph Rowntree Foundation, BT and the Welsh Assembly. 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. All Census data reported in this paper are Crown Copyright. The BHPS data were obtained from the UK Data Archive (University of Essex).

NOTES

(1) http://java.sun.com/

(2) http://www.sas.com/

(3) http://www.smc.kiruna.se

(4) http://www.jrf.org.uk

(5) http://www.lse.ac.uk/Depts/sage/discussion.htm ; see also Evandrou *et al.* (2001).

(6) It should be noted that the class composition table is actually a subset of the employment table, *i.e.* class is allocated only to households with an economically active head. The three class categories are made up from Socio-Economic Groups (SEGs); the *affluent* group comprises SEGs 1, 2, 3, 4 and 13, the *middle* group is SEGs 5, 8, 9, 12, 14, 16 and 17, and the *poor* group is made up of SEGs 6, 7, 10, 11 and 15.

(7) It should be noted that the city of York as defined here does not include the metropolitan boundaries.

(8) It should also be noted that caution is needed when applying the above algorithm to avoid situations where a 'rare' household would dominate the population of an unusual ward through a relatively high weight. This potential problem may be solved by imposing a maximum limit on a household's weight and may be considered in our future work.

(9) The 2001 Census data are also an obvious benchmark. However, these were not available at the time of the preparation of this paper.

References:

Ballas, D. (2001), *A spatial microsimulation approach to local labour market policy analysis*, unpublished PhD thesis, School of Geography, University of Leeds

Ballas, D., Clarke, G.P. (2000), GIS and microsimulation for local labour market policy analysis*, Computers, Environment and Urban Systems*, *24*, 305-330

Ballas, D. and Clarke, G. P. (2001a), Towards local implications of major job transformations in the city: a spatial microsimulation approach, *Geographical Analysis* 33, 291-311

Ballas, D., Clarke, G.P. (2001b), Modelling the local impacts of national social policies: a spatial microsimulation approach, *Environment and Planning C: Government and Policy, 19,* 587 – 606

Ballas, D., Clarke, G.P., Commins, P. (2001), Spatial microsimulation for rural policy analysis, paper presented at the 41st *European Regional Science Association (ERSA) Congress*, Zagreb, Croatia, August 2001

Ballas, D., Clarke, G.P., Turton, I. (1999), *Exploring Microsimulation methodologies for the estimation of household attributes,* paper presented at the 4th International Conference on GeoComputation, Fredericksburg, Virginia, USA, 25-28 July 1999, copy available from the authors, School of Geography, University of Leeds, Leeds LS2 9JT

Ballas, D, Rossiter, D, Thomas, B, Clarke, G P and Dorling D (2004a), *Geography matters: simulating the local impacts of national social policies*, Joseph Rowntree Foundation, York (forthcoming)

Ballas, D. Clarke, G.P., Wiemers, E. (2004b), Building a dynamic spatial microsimulation model for Ireland, Applied Population and Policy (forthcoming)

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 using Synthesis, *Regional Studies,* 23, 535-548

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, 10-27

Brimblecombe, N., Dorling, D. and Shaw, M. (2000) Migration and geographical inequalities in health in Britain: an exploration of the lifetime socio-economic characteristics of migrants, Social Science and Medicine, 50, 6, 861-878.

Caldwell, S.B., Keister, L.A. (1996), Wealth in America: family stock ownership and accumulation, 1960–1995, in G.P. Clarke (ed.) *Microsimulation for Urban and Regional Policy Analysis,* Pion, London, 88-116

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

Clarke, G.P. (1996), Microsimulation: an introduction, in G.P. Clarke (ed.) *Microsimulation for Urban and Regional Policy Analysis,* Pion, London, 1-9

Clarke, G.P. (ed.) (1996), *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., Fieldhouse, E., Holdsworth, C. (2000), *Analyzing Census Microdata*, Arnold, London

Dorling, D. (1994) The negative equity map of Britain, *Area*, 26, 4, 327-342

Dorling, D. and Woodward, R. (1996) *Social polarisation 1971-1991: a micro-geographical analysis of Britain*, monograph in the Progress in Planning series, 45, 2, 67-122

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

Hancock, R., Sutherland, H. (eds) (1992)*, Microsimulation models for public policy analysis: new frontiers*, Suntory-Toyota International Centre for Economics and Related Disciplines – LSE, London

Hancock, R., Mallender, J., Pudney, S. (1992), Constructing a computer model for simulating the future distribution of pensioners' incomes for Great Britain, in R. Hancock, H. Sutherland (eds) *Microsimulation models for public policy analysis: new frontiers*, Suntory-Toyota International Centre for Economics and Related Disciplines – LSE, London, 33-66

Holm E, Lindgren U, Makila K, Malmberg G (1996), Simulating an entire nation, in G.P. Clarke (ed.), *Microsimulation for Urban and Regional Policy Analysis*, Pion, London, 164-186

Hooimeijer, P (1996), A life-course approach to urban dynamics: state of the art in and research design for the Netherlands, in G.P.Clarke (ed) *Microsimulation for urban and regional analysis*, Pion, London, pp 28-63

Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. Carnegie Inst. Tech. Res. Mem. No. 52.
Marsh, C. (1993), The sample of anonymised records, in A. Dale, C. Marsh (eds) *The 1991 Census Users's Guide*, London, HMSO, 295-311
Orcutt, G.H., Mertz J., Quinke, H. (eds) (1986), *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam
Johnston, R., Pattie, C., Rossiter, D., Dorling, D., Tunstall, H. and MacAllister, I., (1998), Anatomy of a Labour landslide: the constituency system and the 1997 general election, *Parliamentary Affairs*, 51, 2, 131-148
Krupp, H. (1986), Potential and limitations of microsimulation models, in G.H. Orcutt, J. Mertz, H. Quinke (eds) *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam, 31-41
Marsh, C., Teague, A. (1992), Samples of anonymised records from the 1991 Census, *Population Trends*, *69,* 17-26
Mertz, J. (1991), Microsimulation - A survey of principles developments and applications, *International Journal of Forecasting, 7,* 77-104
Middleton, E. (1995), Samples of Anonymized Records, in S. Openshaw (ed.) *Census Users' Handbook*, GeoInformation International, London, 337-362
Mitchell, R., Dorling, D., Shaw, M. (2000) *Inequalities in life and death: what if Britain were more equal?* The Policy Press, Bristol
Sutherland, H., Piachaud, D. (2001), Reducing child poverty in Britain: an assessment of government policy 1997-2001, *The Economic Journal, 111,* 85-101
Taylor, M F, Brice J., Buck, N., Prentice-Lane, E (2001) *British Household Panel Survey User ManualVolume A: Introduction, Technical Report and Appendices.* Colchester: University of Essex.
Vencatasawmy, C.P., Holm, E., Rephann, T. *et al.* (1999), *Building a spatial microsimulation model*, paper presented at the 11th Theoretical and Quantitative Geography European colloquium, Durham Castle, Durham, 3-7 September 1999
Voas, D. W. and P. Williamson (2000) An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata, *International Journal of Population Geography, 6, 349-66* 6, 349-66.
Wiemers, E., Ballas, D, Clarke, G P (2003), *A Spatial Microsimulation Model for Rural Ireland—Evidence from the 2002 Irish Census of Population*, paper to be presented at the Annual Meeting of the Population Association of America (PAA), Minneapolis, Minnesota, USA, 1-3 May 2003
Williamson, P. (1992), *Community care policies for the elderly: a microsimulation approach,* unpublished PhD Thesis, School of Geography, University of Leeds
Williamson, P. (1996), Community care policies for the elderly, 1981 and 1991: a microsimulation approach, in G.P. Clarke (ed.) *Microsimulation for Urban and Regional Policy Analysis*, Pion, London, 64-87
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,* 785-816