A Synthetic Demographic Model of the UK Population: Methods, Progress and Problems

Mark Birkin, Andy Turner and Belinda Wu,
School of Geography, University of Leeds

1. Project Objectives

MoSeS (Modelling and Simulation for e-Social Science) is a research node of the National Centre for e-Social Science. The aim of the project is to develop a national demographic simulation which is specified at the level of individuals and households. Such a model can form the basis for a wide range of applications in both research and public policy analysis.

The simulation model has four distinct elements: a baseline demographic model; a dynamic forecasting model; a simulator for social and economic activities; and a set of scenario-based policy modules. This paper describes progress to date in implementing the baseline demographic model. In Section 2 of the paper we will discuss and evaluate the techniques which we have used to generate population baselines. We then review the e-Science requirements of our project, and comment on progress to date in establishing an appropriate infrastructure. The final section of the paper places this work within the context of the broader objectives of the project, including future directions and plans.

2. Population Recreation

2.1 Architecture

The overall architecture of the baseline model is summarised in Figure 1. The objective of the process is to create a database of individuals for each output area (OA).[1] These individuals are drawn from the Individual Sample of Anonymised Records (ISAR). As the name suggests, the ISAR is an incomplete subset of census records (3%) which have been anonymised through the removal of any spatial referencing below the level of ten standard regions. The simulation extracts individuals from the ISAR with a view to creating a population in each output area which is maximally consistent with Census Area Statistics (CAS) of that OA. However it must also be recognised that there are many CAS cross-tabulations for each census output area; and that these have themselves been subject to various random adjustments for the further protection of individual confidentiality (see Rees et al, 2002, for more detail). Therefore the concept of 'maximally consistent' is not necessarily either unique or easy to define.

In order to address this problem, we have categorised the CAS used into three types - constraining, optimising, and co-varying. Attributes which are included in the constraining set should be perfectly reproduced in the simulation. For example, if we constrain by AgeGroup on the Household Reference Person (HRP), then, if there are 10%

[1] An output area is a geographical unit from the 2001 census comprising around 100 households.

of HRP in AgeGroup 75 and over in the census datasets, then 10% of households in the simulation will be created with HRP in AgeGroup 75 and over. Where an attribute is included in the optimising set, then the algorithm will be encouraged to produce selections which approach these criteria. For example, if car ownership were an optimisation criterion, and average car ownership in an OA is 2 cars per household, then a solution with 2.02 cars per household would be adjudged as superior to another with 2.05 cars per household. Attributes within the co-varying set are neither constrained nor optimised, but we assume that these attributes will co-vary with other attributes which are so treated. For example, if social class is strongly correlated with car ownership, and car ownership is optimised, then we would expect to get realistic social class distributions as a 'by-product' of good car ownership distributions.

2.2 Issues

Three major issues to be addressed in the modelling process concern the estimation method, the deployment of a heuristic, and the source of individual data.

a) Estimation. Many microsimulation models have used the principle of synthetic estimation. A model is used to generate individual characteristics on a progressive basis using compound probabilities. Suppose that an individual aged 55 works as a printer lives in Armley with a wife and no other dependents. What is the probability that such an individual suffers from long-term limiting illness? Such probabilities could be extracted from census data and used as a basis for assigning health status to our synthetic individual[2] (see for example Birkin and Clarke, 1988). The current model uses the principle of 'reweighting': we already have a distribution of individuals (within the ISAR) but these are not representative of each OA. So we need to select from the ISAR in order to define a subset which is more representative. This can be thought of as assigning weights of one or zero to each ISAR reecord.

b) Heuristic. The synthetic models described above do not require heuristics, although complex algorithms often involving multi-dimensional iterative proportional fitting are often required for the estimation of probabilities. In order to reweight data, the most common heuristic is simulated annealing, which has been found to perform well for problems of this type (e.g. Stillwell M Birkin and M Clarke (1988) SYNTHESIS: A SYNTHetic Spatial Information System for urban modelling and spatial planning. *Environment and Planning A, 20,*1645-1671 et al, 2005). In this research, we experiment with genetic algorithms (GAs) as an alternative. Intuition suggests that a GA should be well-suited to an optimisation in which zero-one weights are to be applied to a database (a natural gene string). One would expect the main drawback of such an approach to be its computational intensity, but equally, the process is easily paralellized and can be readily supported within e-Science by high performance computing resources.

[2] Sampling would be typically be conducted using Monte Carlo procedures. If there is a 50% chance of illness, then we would extract a random number between 0 and 1. If this number is less than 0.5, the individual is assumed to be well, otherwise ill.

c) Data. The ideal source of individual data for the model would probably be the 2001 census Household SAR, since it contains data not just about individuals but also the way in which these are joined within households. Unfortunately the Household SAR has only recently become available, and the licensing conditions are somewhat more restrictive than those of the ISAR. Our simulations to date have therefore used the ISAR as a base with household formation considered as a separate modelling task.

2.3 Evaluation of our performance

Detailed testing of the model to date has been confined to the Leeds Local Authority District (LAD), for reasons which will be elaborated in Section 3 below. Our main model performance criterion has been to produce spatial distributions of a series of demographic attributes, and to compare modelled data to known distributions. These comparisons have been performed at a ward level within the Leeds LAD using an Index of Dissimilarity. Results are shown at Table 1.

The major themes emerging from Table 1 would appear to include the following:

i. The relationship between constraining, optimising and co-varying attributes is as expected, with closest adherence for constraints and the loosest for co-variation;
ii. Even the constraining attributes are not distributed perfectly, which could be a facet of some of the data issues described earlier;
iii. Co-variation does not appear to be very effectively represented within the algorithm at present. For example, ethnic status is not as strongly spatial clustered within the model as in reality. This probably reflects the fact that none of the constraining or optimising attributes currently selected is closely related to ethnic status. An interesting question is what constraining and optimising attributes might be selected in order to give the best overall profile of a city and its constituent neighbourhoods. The results from further tests of this type will be reported in the full paper.

3. e-Science & architecture

3.1 Parallelism

It was hinted earlier that the GA procedure is computationally intensive. At the time of writing, the method has been implemented on a 32-node Beowulf cluster in the School of Geography at the University of Leeds, using mpj protocols for data transfer and load balancing within a java application (Baker et al, 2004). This cluster is currently able to generate solutions at a rate of approximately 8,000 OAs per day. In other words, it will take around one month to recreate the UK population! This is not necessarily a problem, although it does appear likely that several iterations would be necessary before an acceptable base solution can be determined. The next logical steps are both to enhance

the efficiency of the algorithm, and to utilise the more powerful processing capabilities of the national e-Science infrastructure, specifically the White Rose Grid.

3.2 Data

It was shown earlier that population creation relies on a number of data sources, specifically the CAS and the ISAR. At a later stage in the development cycle, we envisage the inclusion of further data sets such as HM Land Registry or British Household Panel Survey (BHPS). This presents obvious problems relating particularly to virtualisation and access control. In respect of virtualisation, the hope is that the major census datasets may be grid-enabled in order to support easy access from applications such as MoSeS. Specific proposals involving Leeds, Manchester (as data hosts) and others are currently under discussion. Regarding controlled access to data, we are considering a number of protocols which might be adopted by MoSeS. For example, an elegant solution would be to build an explicit certification linkage between MoSeS and Athens which establishes that users have the necessary permissions to access the underlying data. An inelegant solution to the same solution might be to require that all users register independently with MoSeS, and at that time are required to demonstrate permissions to use each of the component databases.

4. Future directions

A number of current and future directions for this research have already been mentioned in the discussion. For example, we need to undertake tests to find out which attributes are best-suited for the problem of constraining and optimisation. It may also be of interest to explore whether different sets of base populations might be deployed for different kinds of applications. For example, does health policy research require a base population which is heavily constrained and optimised by health variables? We also need to generate efficiencies within the current algorithm; and if these are not extremely dramatic, we need to find bigger computational resources for its execution.

At the same time, work is also being undertaken on dynamic algorithms which will be able to take our base population and forecast future changes. New kinds of behavioural and activity variables (such as journey-to-work, shopping trips, or healthy lifestyles) will be introduced in order to support policy applications of the model. We also expect to consider more sophisticated interactions between individuals, and also between individuals and their environment: for example, the influence of social networks on demographics and economic activity.

MA Baker, H Ong, A Shafi (2004) A Status Report: Early Experiences with the implementation of a Message Passing System using Java NIO, Research Report, University of Portsmouth, at http://dsg.port.ac.uk/~shafia/res/papers/DSG_2.pdf

M Birkin and M Clarke (1988) SYNTHESIS: A SYNTHetic Spatial Information System for urban modelling and spatial planning. *Environment and Planning A, 20,*1645-1671

P Rees, D Martin, P Williamson (2002) Census Data Resources in the United Kingdom, in P Rees, D Martin, P Williamson (eds) *The Census Data System*, Wiley, London.

J Stillwell, M Birkin, D Ballas, R Kingston, P Gibson, P (2004) Simulating the City and Alternative Futures, in Unsworth, R., and Stillwell, J. (eds) *Twenty-first century Leeds: Geographies of a Regional City*, Leeds University Press.

Table 1

Attribute	IoD	Type	Attribute	IoD	Type	Attribute	IoD	Type
White	0.04	V	Semi-Detached	0.11	O	Unemployed	0.20	V
Males	0.04	V	Public Transport	0.11	V	rented	0.21	O
full time	0.06	V	Married	0.11	V	Single	0.21	V
25-44	0.06	C	manual	0.11	V	16-24	0.22	C
part time	0.06	V	Own Vehicle	0.12	V	Flats	0.26	O
65+	0.07	C	No Qual's	0.13	V	Detached	0.29	O
Long-term ill	0.07	O	OwnerOccupied	0.14	O	Mixed	0.29	V
45-64	0.07	C	Professional	0.14	V	Level3	0.30	V
Co-habiting	0.07	V	Walk	0.16	V	Other	0.41	V
intermediate	0.08	V	AveCarOwnership	0.17	V	Asian	0.46	V
Level2	0.09	V	Terraced	0.17	O	Students	0.46	V
Level1	0.09	V				Black	0.52	V
Under16's	0.09	C						

Note:
Index of Dissimilarity (IoD) is an index between 0 (perfect correspondence) and 1 (no correspondence)
Attribute types are constrained (C), optimised (O) and covarying (V). For discussion, see text.