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

Mark Birkin[1], Andy Turner[1] and Belinda Wu[1]

[1]School of Geography, University of Leeds

Abstract. The paper reports on progress with the development of a Population Reconstruction Model which makes a synthetic representation of the entire UK population and its constituent households. A prototype version which deploys a genetic algorithm is evaluated against alternative methods, and recommendations for further development of the method are discussed.

1. Project Objectives

MoSeS (Modelling and Simulation for e-Social Science) is a research node of the National Centre for e-Social Science (NCeSS). 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 Background

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 and lives in Armley with a wife and no other dependents. What is the probability that such an individual suffers from limiting long-term

illness (LLTI)? Such probabilities could be extracted from census data and used as a basis for assigning health status to our synthetic individual[1] (see for example Birkin and Clarke, 1988).

The probabilities associated with direct synthetic estimation are often established using the techniques of Iterative Proportional Fitting (Fienberg, 1977). IPF allows that multiple census attributes can be compounded into a single table. In the example above, it might be the case that LLTI can be tabulated against age, or against marital status, or against occupation. We might also know age by marital status, and age by occupation. IPF would be used to estimate the full distribution of LLTI by age by marital status by occupation. An algorithm to support Iterative Proportional Fitting of census data with large number of variables is described by (Rees et al, 2005).

The Moses PRM 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 record. An approach to reweighting using IPF based on the American PUMS (Public Use Micro Sample, broadly equivalent to the UK SARs) is described by Beckmann et al (1996). As in the UK, the problem with the PUMS is that it is spatially referenced only to supertract level, with areas of about 100,000 individuals (Beckmann et al, 1996, page 415). The authors identify four key attributes from the main US census tables – age, household composition, ethnicity and income. These attributes are combined for each small area (census tract) using IPF. Individual records are then selected from the PUMS in accordance with the combined distribution. This method is not directly applicable to the UK context for two reasons: firstly that the spatial coding within the UK SARs is even less discriminating than the PUMS[2]; and secondly that the key income variable is not captured in the UK census.

In order to reweight data from the UK SARs, the most common heuristic has been simulated annealing e.g. Williamson et al (1998), Ballas and Clarke (2004). The performance of this method has been good, although it has not been benchmarked explicitly against other approaches. The major drawback of simulated annealing is that it requires significant computation, particularly in the context of a national simulation involving more than two hundred thousand small areas. The method can potentially be made parallel across areas but not within areas. In other words, within a multi-processor environment, we could send data for different areas to a number of different processors, but it would make no sense to send data for a single area to more than one processor.

An alternative method to simulated annealing is a genetic algorithm, as suggested by Williamson et al (1998). 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 this process is easily parallelised both within and between areas. In other words, the solution method will involve the creation and evolution of a candidate population of solutions for each area, and there is no reason why each candidate solution need not be assigned to a different processor within a multi-processor environment. The results of some experiments with a GA implementation of the PRM are reported in this paper.

---

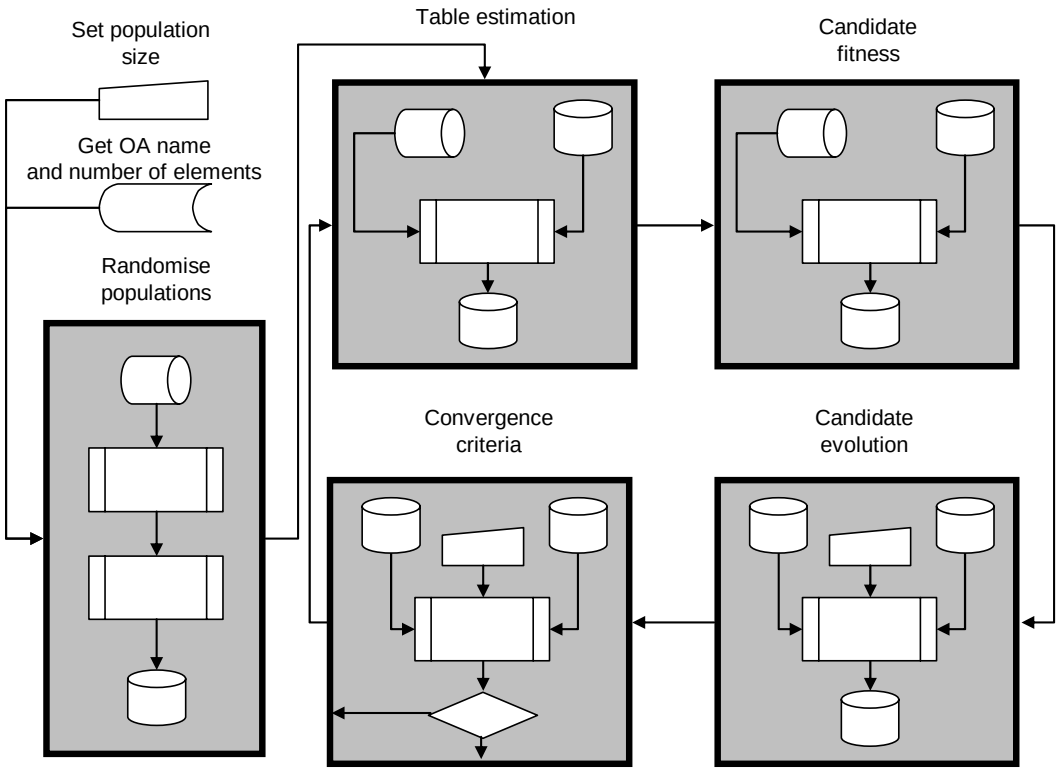
[1] Sampling would typically be conducted using Monte Carlo procedures. If there is a 50% chance of illness, then we would extract a random number between 0 and 1. If this number is less than 0.5, the individual is assumed to be well, otherwise ill.

[2] The UK HSAR is referenced only to the level of Local Authority Districts, which can have up to a million residents. The ISAR is even less detailed with coding for 13 standard regions.

2.2 Architecture

The overall architecture of the baseline Population Reconstruction Model (PRM) is summarised in Figure 1. The objective of the process is to create a database of individuals for each output area (OA).[3] 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 detailed spatial referencing.

Figure 1. Architecture of the Population Reconstruction Model



Once individuals have been extracted for each output area, they need to be assembled into households. Both individual and household composition are important in the characterisation of small areas. To take a simple example, student areas have a large number of individuals who may share a single dwelling unit, but are treated as separate households. In multi-ethnic areas, it is much more likely that multiple individuals will be connected within a single extended family household. At the second stage in the algorithm, individuals with appropriate characteristics are selected as a Household Representative Person (HRP). According to the demographic characteristics of the HRP, further individuals my be added to the household (for example, wives and children may be added to HRPs who are married or cohabiting).

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

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

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 - controls, optimisations, and co-variants. Attributes which are included in the control set should be perfectly reproduced in the simulation. For example, if we control by AgeGroup on the Household Reference Person (HRP), then, if there are 10% of HRPs 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-variant set are neither controlled 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.

2.3 Evaluation of model 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 I.

Table I Results from the PRM (Version 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.

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

i. The relationship between controls, optimising and co-varying attributes is as expected, with closest adherence for constraints and the loosest for co-variation;
ii. Even the control 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 spatially clustered within the model as in reality. This probably reflects the fact that none of the controls or

optimising attributes currently selected is closely related to ethnic status. An interesting question is what controls and optimising attributes might be selected in order to give the best overall profile of a city and its constituent neighbourhoods.

2.4 Issues

The experiments which have been reported in the previous section have allowed us to draw a number of conclusions which can be introduced to the next version of the algorithm.

2.4.1 Households and communal establishments

The ISARs include individuals from both private households and communal establishments. A communal establishment is essentially an institution with more than twelve residents, such as a boarding school, jail or institute for young offenders. Both private households and communal establishments need to be incorporated within the model. Communal establishments are likely to be of particular importance in health applications e.g. old people's homes, nursing homes, and so on.

At present, however, there is a confusion between the two household types within the model, as individuals in private households can be selected as candidates for communal establishments, and vice versa. Private households and communal establishments need to be treated as two different sub-populations for the PRM. Private households are represented within the Household SAR. Members of communal establishments are represented within the Individual SAR. Separate tables exist within the census small area statistics for private households and communal establishments.

Similar procedures can be used in the reconstruction of communal establishments as for private households i.e. they can be treated as 'big households'. This is important as most previous research focuses solely on private households. In the first instance, however, we will concentrate purely on private households.

2.4.2 Nature of the constraints

The existence of two types of constraining totals has already been noted. **Control totals** are irrevocably binding. For example, if there is a count of 10 HRPs aged between 20 and 29 in the small area statistics, then there would be exactly 10 HRPs aged between 20 and 29 in the PRM for the same area. **Optimisation totals** are counts which are used to steer the fitness of the solution within the GA. If economic activity is a type 2 constraint and there are 10 economically active HRPs in a small area, then we might expect the solution to contain 9 EA HRPs (a good solution) or 6 EA HRPs (a poor solution), but it does not necessarily contain 10 EA HRPs.

Within Version 1 of the model, control totals are applied to the age, sex and marital status of HRPs[4]. Optimisation totals are applied to economic activity, household composition, individual demographics and health status[5].

[4] Control totals are derived from CAS002 (individual age) and from CAS020 (household size).

[5] Economic activity (unemployment, retired, permanently sick or disabled, economically inactive – from KS9B and KS9C); household composition (children, dependents, lone parents - KS020); age, sex and marital status (of individuals - CAS001, CAS002); health status & limiting long-term illness (CAS008)

Whether this is the correct list of attributes to use, and whether it is sufficiently extensive is a question for ongoing research (cf Beckmann et al, 1996 and Section 2.3 above).

2.4.3 Household SAR versus Individual SAR

Version 1 of the PRM has been implemented using the ISAR for practical reasons, as the HSAR has only been available to us since early in 2006. There are arguments for continuing with the ISAR as the microdata base for the Moses project. It is larger than the HSAR, and provides a regional geography. It covers the communal establishment population as well as the household population; and it covers the whole of the United Kingdom. However the big difficulty with the ISAR is the difficulty in reconstructing household relationships, to get the right composition of family members within each household.

The HSAR provides bundles of people in households, so that relationships and membership come from real data. This is a crucial advantage, regardless of the problem that the HSAR only covers England and Wales, not Scotland and Northern Ireland and has no regional indicator; and that the HSAR does not cover the communal establishment population.

It has therefore been resolved that the Household SAR (HSAR) needs to be used as the basis for the reconstruction of private households. The Individual SAR (ISAR) will be used as the basis for the reconstruction of communal establishments. The rationale for this is that individuals within communal establishments are captured within the ISAR, but not within the HSAR. This makes logical sense given that communal establishments are collections of individuals rather than households, although it is inconvenient to the extent that we need to consider two sets of data rather than just one.

We have no definitive answer at this stage as to how the PRM might be later be extended to include Scotland and Northern Ireland. Some kind of synthetic generation of a household sample based on the individual records would appear to be indicated.

2.4.4 Use of larger zones than output areas

The application of the PRM for output areas is computationally challenging to the extent that there are over 200,000 of these areas, and the current implementation of the PRM uses several weeks of parallel processing to recreate the population. This implementation is discussed further at Section 4 below. Nevertheless the continued use of output areas is questionable in view of issues relating to data issues and model robustness.

The data issue concerns the small cell adjustment method (SCAM) which means that many small area tables are potentially inconsistent when considered at the level of individual households. The SCAM procedures are applied to maintain confidentiality within output areas, and in practice they mean that counts of 1 or 2 can never be observed within the Census Area Statistics. Such counts are either rounded down to zero, or upwards to 3. Because this rounding is applied to aggregate totals, it is possible to generate inconsistencies between census tables e.g. the count of adults aged between 30 and 35 might appear to be different between two different tables.

The robustness issue arises from the problem of generating reliable model estimates from small area populations. Suppose that we wish to generate populations by single year of age (which is essential in a dynamic microsimulation) and have an age range of 0 to 100+ (to monitor old age issues), we will have average age-sex populations of 1.3 within each output area. This looks hopelessly unrealistic for modeling disease incidences or mortality probabilities. Recent ONS work in relation to mortality indicators suggests a minimum area size in the order of 5,000

residents. We have therefore settled on the MSOA (Middle Level Super Output Area) geography – 7,193 areas in England and Wales with an average population of 7,300.
Note that at some stage within the modelling procedure, the spatial coding does need to be refined to something like OA level. This is necessary for flexible aggregation and detailed model analysis: for example, MSOAs will be much too large for analysis of service provision. We assume that this is a discrete step which follows the main part of the PRM model.

2.4.5 Dynamic simulation

The adoption of the HSAR within private households allows the procedures for dynamic simulation and population reconstruction to be separated to the maximum extent possible. The reconstruction of the population base is essentially a once-and-for-all process. It is unclear at the present time whether multiple baselines might be advantageous e.g. for bootstrapping purposes. The dynamic simulation, on the other hand, may be run for multiple scenarios, or customised for the purpose of a particular planning exercise. The particulars of the dynamic simulation model will be reported at a later date.

2.4.6 Algorithm

As discussed above, the modelling process has been constructed around a genetic algorithm. Individual solution populations are represented as a string of values of length N_i, where N_i is the number of people (or households) in small area i. Each element of the string is a pointer to a record in either the HSAR or ISAR. The populations are all evolved in accordance with the usual principles of mutation and recombination within genetic algorithms. The precise nature of the algorithm, for example population size, mutation frequency, evaluation of fitness and so on, are all subject to further investigation.

The possibility that GAs are not an ideal solution mechanism for the PRM can not yet be discounted. In other words, that the relatively poor results from Version 1 of the model are in some measure related to the performance of the algorithm itself. Some experiments comparing the GA to Iterative Proportional Fitting are reported in Section 3 of the paper.

3. Model Benchmarking

In this part of the paper, we report on some experiments with a genetic algorithm. These results are benchmarked against an estimation method which is inspired by IPF.

3.1 Description of the GA

The test GA has the following characteristics:

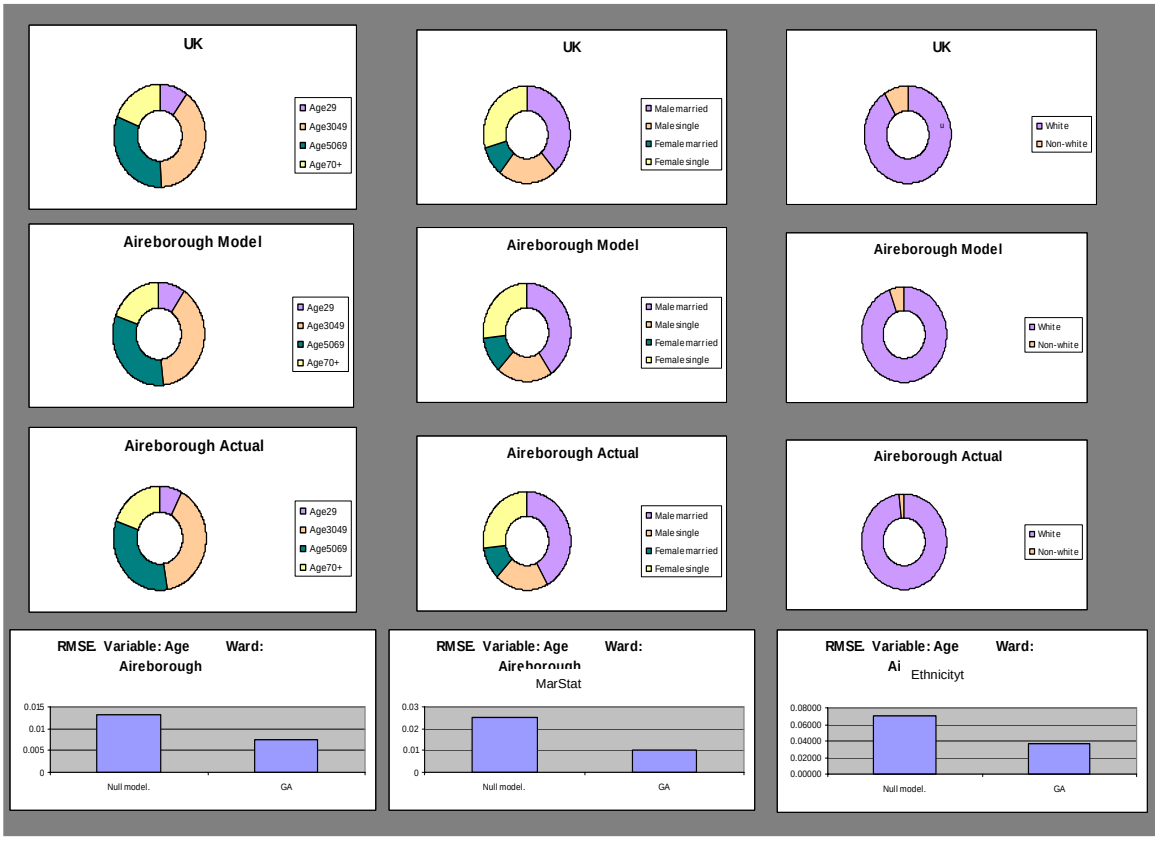
- it is implemented on a ward-by-ward basis
- for each ward, a random population is created. The random population consists of an extract from the HSAR. N_i households are selected. Duplicates from the HSAR are not permitted.
- The 'fitness' of each solution is evaluated against two small area census tables – age, sex and marital status of HRP (CAS003), and ethnicity of individuals (KS006). The fitness value is an unweighted sum of the total absolute error between the cells of the two census tables, as observed and 'predicted' within each random population.

- New solutions are 'bred' from one iteration to the next. Parents are selected at random, but in accordance with fitness such that good solutions are more likely to become parents of a new solution. From each pair of parents, two new solutions are bred. One is the mirror image of the other. In other words, there is a 'crossover' operation such that a part of each parent solution is exchanged with the other in order to create two 'child' solutions.
- The best performing ('fittest') solution is always retained from one iteration to the next. 'Mutations' are occasionally performed, in which a member of the solution population is exchanged at random with another household from the HSAR.
- The size of the gene pool is arbitrarily selected to be 50 solutions. The algorithm runs for 100 complete iterations. Practical tests to date have demonstrated that by this time the solutions have become uniform, with the exception of random evolutionary mutations.

3.2 Results from the GA

The first implementation of the GA is for the Aireborough ward in Leeds MD. This area was selected as first in the alphabetical list of Leeds wards. Results from this model for three variables of age, marital status and ethnicity are shown in Figure 2.

Figure 2. Experimental GA Results for the Aireborough ward



Three sets of distributions are shown in Figure 2. These are the distributions for Aireborough (from the UK census) [*Aireborough Actual*], the distributions for the UK (from the census) [*UK*], and the distributions predicted by the GA (after 100 iterations) [*Aireborough Model*]. The UK

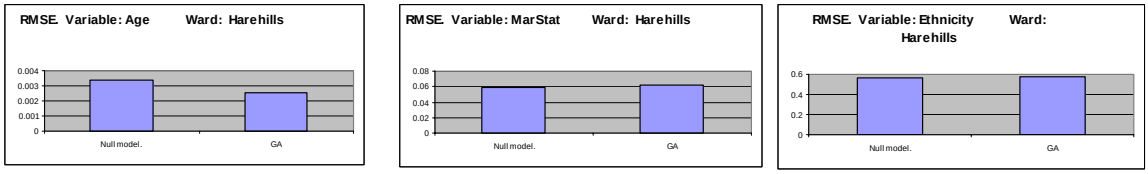
distribution is treated as a null model. This is the model which we would expect to be given by a random extract from the HSAR. We are interested in the comparison between the GA and the null model. The comparison is based on the root mean squared error (RMSE) between two distributions. For each variable (age/ marital status/ ethnicity) we calculate the RMSE between the null model and the small area distribution, and between the GA model and the small area distribution:

$$RMSE_i^m = \sqrt{\sum_{k=1}^{N^m} (x_i^{km} - \hat{x}_i^{km})^2} / N^m$$

Where x_i^{km} is the proportion of the population in area i which falls in class k for attribute m; $\hat{x}_i^{km}$ is a model approximation (either GA or null); and N^m is the number of classes for attribute m.

The RMSE for the null model and GA are shown at the bottom of Figure 2. The results indicate that the GA gets about half way from the null solution to the actual distribution on each of the attributes. Equivalent results for a second ward – Harehills – are shown in Figure 3. Harehills has been selected because this area is very different to Aireborough. In particular, it is much more ethnically diverse. The results for Harehills are much less satisfactory. They show some improvement in the age mix between the null model and the GA, but no improvement for marital status and a marginal deterioration for ethnicity. The intuitive explanation for these results is that ethnic groups are a fairly small minority of the national population. It is difficult for the GA to find enough of these minority populations to evolve towards a satisfactory solution.

Figure 3. Experimental GA results for Harehills



3.3 Iterative Proportional Sampling

In addition to its poor performance, the test algorithm of Section 3.1 is also quite inefficient. We have therefore decided to implement a much simpler algorithm in order to provide benchmarks for the performance of the PRM. The idea behind this algorithm is that the population of a small area can be constructed by sampling from the HSAR. The individual selection probabilities can be weighted according to the characteristics of each household. Is it possible to identify appropriate weights which allow the population of individual small areas to be generated with reasonable accuracy.

The algorithm therefore looks as follows:

- Draw a random population from the HSAR
- Construct cross-tabulations from the synthetic population, and compare these with the actual populations for a small area

- Adjust weights upwards for attributes that are underrepresented, and adjust weights downwards for weights that are overrepresented (for example, in an inner city multi-ethnic area like Harehills, increase the probability of selecting a non-white household)
- Reconstruct the cross-tabulations. Keep adjusting the weights until they stabilise.

The algorithm has been implemented using four sets of univariate census tables – age of HRP (CAS003), ethnicity (KS006), socio-economic status (KS012) and health status (KS008). Some results for Aireborough are shown in Figure 4, and a series of comparisons between IPS and the null model are shown in Table II.

Figure 4. Experimental IPS results for Aireborough

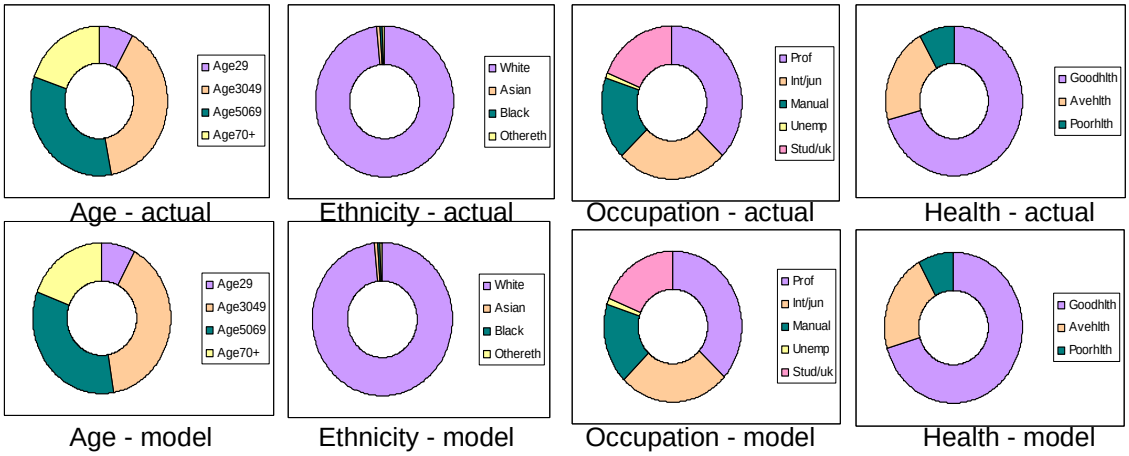


Table II. IPS Results for two covarying attributes: Leeds wards

RMSE	IPS model				Null model			
	Age	Ethnicity	S-e status	Health	Age	Ethnicity	S-e status	Health
Aireboroug	0.002141	0.000254	0.002614	0.003733	0.064157	0.077753	0.085772	0.036714
Armley	0.00656	0.001531	0.007372	0.022691	0.041474	0.010876	0.073748	0.029791
Barwick an	0.003219	0.000594	0.005411	0.013231	0.066295	0.084221	0.060583	0.018433
Beeston	0.001865	0.000957	0.002509	0.00451	0.017786	0.045511	0.091817	0.039333
Bramley	0.003863	0.002355	0.005136	0.000909	0.025545	0.064818	0.082868	0.032823
Burmantofl	0.00423	0.001653	0.005097	0.011888	0.027207	0.039118	0.135752	0.083432
Chapel Alle	0.002677	0.023809	0.002498	0.003355	0.051944	0.301776	0.054562	0.032715
City and H	0.001872	0.003927	0.002163	0.008116	0.080564	0.099689	0.115753	0.065235
Cookridge	0.004065	0.004517	0.007346	0.00648	0.08842	0.030693	0.106293	0.030947
Garforth ar	0.005181	0.0009	0.003748	0.010319	0.082352	0.087092	0.061939	0.003784
Halton	0.005894	0.000624	0.009736	0.020928	0.088035	0.082871	0.070423	0.014854
Harehills	0.003551	0.005945	0.004751	0.001619	0.079142	0.404974	0.112688	0.02289
Headingley	0.0078	0.004206	0.005959	0.006719	0.421222	0.07694	0.297947	0.10579
Horsforth	0.003586	0.003053	0.003004	0.002476	0.066528	0.072125	0.134919	0.060029
Hunslet	0.004224	0.000867	0.005515	0.011537	0.020741	0.070044	0.153888	0.063594
Kirkstall	0.008532	0.00276	0.007713	0.019204	0.119523	0.029027	0.036436	0.004044
Middleton	0.002058	0.002637	0.003758	0.004791	0.055792	0.078372	0.082601	0.01738
Moortown	0.009618	0.005005	0.015904	0.03951	0.063019	0.115566	0.117527	0.009635
Morley Nor	0.006991	0.001255	0.007527	0.012665	0.050938	0.073031	0.079805	0.022847
Morley Sou	0.009891	0.00177	0.00954	0.024848	0.04586	0.069835	0.076671	0.024807
North	0.003715	0.001474	0.002016	0.003129	0.074873	0.032951	0.146375	0.025406
Otley and \	0.003588	0.000541	0.007236	0.017461	0.086433	0.079429	0.108945	0.048594
Pudsey No	0.007309	0.004351	0.006623	0.021738	0.045728	0.026515	0.084015	0.028843
Pudsey So	0.004534	0.000733	0.007046	0.009073	0.04782	0.063913	0.048016	0.003825
Richmond	0.006007	0.001356	0.00796	0.01653	0.015841	0.056237	0.137448	0.072075
Rothwell	0.005425	0.002844	0.008589	0.012911	0.056036	0.081815	0.040404	0.005961
Roundhay	0.019788	0.007764	0.020659	0.042093	0.058582	0.09722	0.190523	0.058486
Seacroft	0.014419	0.000934	0.018715	0.047329	0.026822	0.071831	0.152776	0.071824
University	0.002098	0.000689	0.004414	0.011448	0.247322	0.209332	0.231973	0.013753
Weetwood	0.010191	0.001679	0.01049	0.035195	0.098377	0.011537	0.141716	0.0383
Wetherby	0.002955	0.001408	0.003655	0.007536	0.123048	0.078985	0.137655	0.066137
Whinmoor	0.008673	0.002379	0.011763	0.032568	0.065871	0.056598	0.066082	0.03268
Wortley	0.001886	0.001763	0.00252	0.007443	0.021136	0.064277	0.089167	0.023143
	0.005709	0.002925	0.006939	0.014969	0.076498	0.086211	0.109306	0.036609

Coincident distributions have been assessed using two further distributions which are car ownership (KS017) and educational attainment (KS013). Detailed analysis of these results shows that the model is clearly effective in explaining some of the variations in car ownership and educational attainment, but detailed analysis also shows that considerable unexplained variation remains. In particular, car ownership levels tend to be lower than predicted in inner city areas, while high levels of educational attainment in student areas like Headingley and Weetwood are not fully accounted within the model. Neither of these results is entirely surprising.

3.4 Assessment of results

The results from the GA model are somewhat disappointing, even allowing that this is very much a prototype. The results can be expected to improve once control totals for essential demographics are introduced. The inclusion of a wider range of optimisation tables should also enhance the performance of the model. Nevertheless it appears at this stage that other improvements to the algorithm may need to be discovered if performance is to become satisfactory. One possibility is that the populations might be sequenced according to some kind of household segmentation or typology, so that individual households are typically exchanged with others of a similar type, rather than something completely different.

Early results from the IPS approach are already much more encouraging. As well as providing a set of benchmarks for algorithm performance, it is worth considering whether IPS might provide an alternative solution mechanism within the PRM. At such time, it would be possible to consider which are the right attributes to incorporate within the simulation, and whether multivariate tabulations might yield better results.

4. e-Science & Architecture

4.1 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 takes several weeks 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.

4.2 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, we hope to exploit progress on the GEMS (Grid Enabling Mimas dataSets) and NCeSS e-Infrastructure projects. 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 problem 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.

5. Future directions

In this paper we have described a genetic algorithm for the synthetic reconstruction of the UK population, and an associated solution architecture. The results from an initial implementation of the algorithm were found to leave considerable room for improvement. A series of adjustments to the algorithm have been described which will yield more satisfactory results in the next phase of our research.

Experimental results have also established benchmark levels of performance for the PRM. These results are suggestive of an alternative strategy for the reconstruction of populations should the GA continue to prove unsuitable for this purpose.

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.

References

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

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

RJ Beckmann, K Baggerly, M McKay (1996) Creating synthetic baseline populations, *Transportation Research A, 30*, 415-429.

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.

P Rees, J Parsons. and P Norman (2005) Making an estimate of the number of people and households for output areas in the 2001 Census. *Population Trends, Winter 2005*.

P Rees, J Stillwell and A Tyler-Jones (2004) The City is the People: Demographic Structure and Dynamics, in Unsworth, R., and Stillwell, J. (eds) *Twenty-first century Leeds: Geographies of a Regional City*, Leeds University Press.