Monograph:

Smith, D.S., Clarke, G.C. and Harland, K.H. (2007) SimHealth: Estimating Small Area Populations Using Deterministic Spatial Microsimulation in Leeds and Bradford. Working Paper. The School of Geography , The University of Leeds.

School of Geography Working Paper 07/6

(Working Paper 07/06)

SimHealth: estimating small area populations using deterministic spatial microsimulation in Leeds and Bradford

Dianna M Smith, Kirk Harland and Graham P Clarke

Version 1.0
July 2007

School of Geography, University of Leeds,
Leeds, LS2 9JT, United Kingdom

Our full contact details are:

Mail address:

School of Geography
University of Leeds
Leeds
LS2 9JT
United Kingdom

Email:

Dianna M. Smith:
d.m.smith04@leeds.ac.uk

Kirk Harland:
k.harland98@leeds.ac.uk

Professor Graham P. Clarke:
gpclarke@leeds.ac.uk

Acknowledgements

This paper is produced as part of research funded by an Overseas Research Studentship, the University of Leeds (Tetley and Lupton Scholarship), and the School of Geography for a postgraduate maintenance grant. Kim Procter, Myles Gould and Mark Birkin were involved in early model discussions.

The 2001 Census data used in this paper are Crown Copyright and are produced by the Office of National Statistics (ONS). All of the Census data was accessed through CASWEB online, provided by the Census Dissemination Unit (CDU) through the Manchester Information and Associated Services (MIMAS), supported by Manchester Computing, University of Manchester.

All of the health data analysed in this paper was from the 2003-4 Health Surveys for England, led by the National Centre for Social Research and University College London. Department of Epidemiology and Public Health. Data is distributed by the UK Data Archive, Colchester, Essex. The 2003 survey was deposited in March 2005 (SN: 5098) and the 2004 dataset in July 2006 (SN: 5439).

Abstract

The increasing prevalence of obesity and type 2 diabetes in recent decades is often cited as a serious public health concern, lowering life expectancy and costing the National Health Service (NHS) billions of pounds each year. However, measuring diabetes prevalence proves challenging; the best estimates are based on the annual Health Survey for England (HSE) and little is currently available at the small area level.

Simulation models are increasingly used in health research to predict future prevalence, cost of treatment, provision of care and the possible outcomes of policy intervention. Previous research shows the relevance of this technique in modelling the outcomes of changes in taxation and child benefit policy, or analysing health inequalities. This paper introduces SimHealth, a small-area diabetes prevalence model for Leeds and Bradford, West Yorkshire created as part of a generic model framework. The process of configuring an optimal spatial microsimulation model, building on earlier research, is detailed with the aim of improving and extending existing simulation models.

Table of Contents

List of Figures

List of Tables

1 Introduction

This paper outlines SimHealth and the experimental runs created and carried out to identify the most robust method of creating a population microdata set of individuals at the output area (OA) level in Leeds and Bradford, West Yorkshire.

Although 'off the shelf' microsimulation software packages are not available, the Flexible Modelling Framework (FMF) is an application framework that has been developed at the University of Leeds to enable the development and integration of modelling systems using a modular approach (Figure 1). Currently the FMF consists of a framework that handles all application level communication and access to databases through a data access layer, and a Spatial Interaction Model (SIM) component (Harland and Stillwell 2007). The 'MicroSim' component is the latest generic social simulation modelling module that has been developed extending the static deterministic micro simulation techniques applied by Ballas et al (2005). The 'MicroSim' component has been configured to run using 2001 Census and 2003 Health Survey for England (HSE) data, producing the SimHealth model configuration.

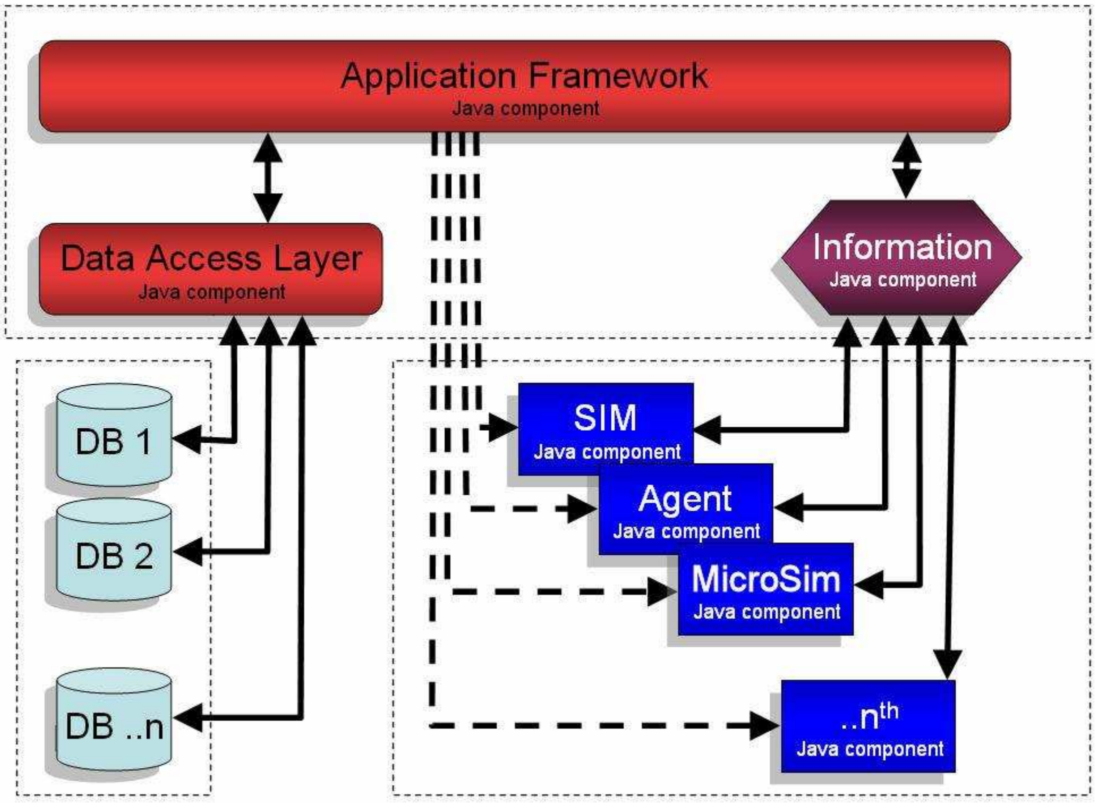

Figure 1. Illustration of the Flexible Modelling Framework

The deterministic method used to create the synthetic populations is a proportional fitting technique, similar to the sample weighting already carried out on the HSE to ensure the results are representative of the general population and adjusted for individual/household refusal. Here, the main model creates these population estimates by constraining the survey population by variables which are not cross-tabulated like the examples provided in Ballas et al. (2005); there are no known relationships between the variables, such as the number of men under age 50 or the total number of non-whites over 15. This model selects individuals from the HSE that most closely match the Census-defined population of an area, reweighting individuals against one constraint at a time. The survey must have a minimum number of variables in common with the Census (typically 3 or 4) in order to carry out the reweighting and provide confidence in the accuracy of resulting estimates; ideally, there should be a strong correlation between the constraint variables and the health outcomes the model is configured to simulate.

One advantage of a deterministic model (such as Ballas et al., 2005) is that the estimated population distributions will be the same each time the model is run. This allows for any number of data changes to be made, with the results from each model being tested against the known population distribution from the Census. If the model and/or the constraint data are changed in any way, the results will indicate the relative success of each change in matching the known (Census-based) population distribution. This characteristic of deterministic models allows us to explore several variations of the initial model and identify the optimal reweighting methodology for a range of health-related applications.

The ideal use of population simulations for health outcomes is to combine the small area prevalence estimates with a policy intervention model to predict the prevalence changes at a local level under proposed policy interventions. The advantage of spatial microsimulation is the usage of detailed survey data to build up the synthetic populations; each person in the simulated population is based on an actual individual in the survey. The 2003 and 2004 HSEs include information on diet, height and weight, waist and hip measurements and diabetes status among over 1600 variables included, allowing us to simulate detailed populations for the entire study area. Using the deterministic reweighting methodology, individuals from the HSE that best fit chosen demographic characteristics (e.g., age, sex, ethnicity, and social grade) from the Census are ‘cloned’ until the population of each small area (initially, an output area of approximately 250 people) is simulated (Table 1). The reliability of these synthetic populations can be validated against other census variables to ensure the synthetic population resembles the actual population (Ballas et al., 2006).

Table 1. Variables from the HSE and Census datasets

VARIABLE	HSE 2003/4 NAME	CENSUS 2001 TABLE/NAME
Age	age	CT003
Sex	sex	CT003
Ethnicity	allcult1/dmethn04	CT003
Social grade	schrpg6	CS066
Marital status	marital	UV007
Tenure	Tenureb	UV043

This reweighting is repeated until each individual has been reweighted to reflect his/her probability of living in each output area. This method ensures that every person has the opportunity to be allocated to every area, however, there may be no 'clones' of an individual in an area, or there may be 150 copies of a single person. The criteria is simply how well each person matches the constraints from the census. The initial baseline model takes each constraint in isolation, so there will be a higher chance that many people will have very small weights; if the dataset were cross-tabulated, there would be fewer individuals selected, but with larger weights (Figure 2, equation 1).

$$n_i = w_i * s_{ij} / m_{ij} \qquad (1)$$

where:

n_i = the new weight of an individual i

w_i = the old weight of an individual i

s_{ij} = is the element s of the small area statistics table for i individual and attribute j

m_{ij} = is the element m of the survey data table for i individual and attribute j.

(above reproduced from Ballas et al., 2005a)

Survey respondents		
ID	**sex**	**Weight (w_i)**
001	male	0.9
002	male	1.2
003	female	0.8

Census output area A (s_{ij} values)
10 males
15 females

Survey totals (m_{ij} values)
8000 males
8500 females

Survey respondents: new weight calculation			
ID	**sex**	**weight**	$w_i * s_{ij} / m_{ij} = n_i$
001	male	0.9	0.9 x 10/8000 = 0.00113
002	male	1.2	1.2 x 10/8000 = 0.0015
003	female	0.8	0.8 x 15/8500 = 0.0014

Figure 2. Worked example of the first part of the reweighting

The reweighting algorithm is very similar to that used by iterative proportional fitting (IPF), although a final process, after the last constraint has been applied, leaves any subsequent iterations of reweighting unnecessary. The sum of all the new weights after the sex constraint is reweighted is calcuated for each OA (NW_o), and should sum to the total population of that OA. Then, the sum of the new weights for all males is calculated (NW_o^c). The ratio of the number of males reweighted by the model (NW_o^c) to the reweighted population for that area (NW_o) is used as a scaling factor on all of the new weights generated by this constraint reweighting process (equation A4). This is needed because the new weights (n_i) are very small values, and would continue to decrease with each further constraint reweighting process as outlined by Ballas et al. (2005a) if the process repeated. Instead, this adjustment brings the weights back to values which are consistent with the real-world population. The step is required because the model is selecting from over 15,000 individuals to estimate a total OA-level population of approximately 250, with each individual who fits an area demographic profile having some ‘share’ in the population of 250. This is the reverse of Ballas’ situation, where a survey population of less than 1000 individuals was reweighted to fit wards which have populations in the thousands. In his work, the initial reweighting is repeated up to twenty times until the new weights converge (2005a). (see equations 2-4):

$$\mathrm{nw}_{\mathrm{o}}^{\mathrm{c}} = \mathrm{oldwt}^{\mathrm{c}} \times \frac{\mathrm{tot}_{\mathrm{o}}^{\mathrm{c}}}{\mathrm{tot}_{\mathrm{s}}^{\mathrm{c}}} \qquad (2)$$

$$\mathrm{NW}_{\mathrm{o}}^{\mathrm{c}} = \sum \mathrm{nw}_{\mathrm{o}}^{\mathrm{c}} \qquad (3)$$

$$\frac{\mathrm{tot}_{\mathrm{o}}^{\mathrm{c}}}{\mathrm{NW}_{\mathrm{o}}^{\mathrm{c}}} \times \mathrm{nw}_{\mathrm{o}}^{\mathrm{c}} \qquad (4)$$

where

nw_{o} is new weight for individual in OA o

NW_{o} is total weight for all individuals in OA o

$^{\mathrm{c}}$ is constraint sub category

$\mathrm{tot}_{\mathrm{o}}^{\mathrm{c}}$ is total population in constraint c in OA o (Census 2001 totals used here)

$\mathrm{tot}_{\mathrm{s}}^{\mathrm{c}}$ is total population in constraint c in survey data (HSE used here)

$\mathrm{oldwt}^{\mathrm{c}}$ is initial starting weight for individual.

Initially SimHealth was intended to reweight the 2003 HSE dataset, however, the 2004 HSE data became available in time for it to be included in the process. This pooling of datasets provides a larger sample from which to build up the synthetic area populations; data aggregation across years was also used in a recent model for obesity at the ward, PCT, and regional level (Moon et al., 2007). Temporal aggregation is not a requirement, as many other researchers use single year data in population prevalence estimates (Pearce et al., 2003) and other microsimulation models (Ballas et al., 2005), however, the inclusion of respondents from two years rather than only one provides a more diverse pool of individuals for the synthetic population.

2 Variable specifications

The correct choice of constraints is vital to building a successful spatial microsimulation model. Each of the constraints must be present in both the base survey (here, the HSE) and the small-area dataset (2001 Census output area tables). The four constraints currently in use (age, sex, ethnicity and social grade) and the two validation variables (marital status and tenure) are all available online (www.casweb.co.uk). The nature of reweighting requires that the variables used as constraints be highly correlated with each other, so correlation analysis was carried out to ensure that the constraints were correlated with the health outcomes and the validation variables.

The reason for choosing the HSE dataset is that it includes many variables of interest to this research. Many of the variables are diet-related, although there is information about various health conditions as well (Table 2).

Table 2. Variables in HSE 2003 relevant to this study

VARIABLE	DESCRIPTION
Cigst1	Cigarette smoking grouped (never, ex-occ, ex-reg, reg)
Porftvg	Grouped portions of fruit and veg eaten yesterday
Fatbanda	Fat score (grouped)
IMD2004	Index of multiple deprivation
diabtype	Type of diabetes
sprtacty	Sport activity level
Shops	Ease of getting to supermarket
Transport	This area has good local transport
Leisure	This area has good leisure things for people like me
Bmivg6	Valid BMI grouped in 6 categories
D7unitg	Units drunk on heaviest day in last 7, grouped

The 2003 HSE dataset (used for the trial models) was cleaned prior to input into the model, with all people who failed to answer one of the constraint questions or other variables of relevance (BMI, diabetes, etc) removed from the dataset. The final test dataset, after this adjustment, included 15,599 respondents. The 2003 and 2004 HSE datasets were pooled to create a more diverse base population planned for use in the final simulations; the initial merged dataset contained 37,021 individual records (respondents). Not all of the variables were consistent between the 2003 and 2004 HSE datasets, although both datasets included questions on diet, BMI, waist/hip ratio, diabetes and physical activity. The merged dataset thus had a total ethnic population created from the 2003 category allcult1 and the 2004 category of dmethn04 (derived ethnicity). The 2004 HSE included a boost sample of minority ethnic groups, however, the general population survey for this year did not include any white respondents with type 2 diabetes. If only the 2004 dataset was included for the reweighting process, then there would not have been any white diabetics included in the final disease estimation. In order to have the disaggregated ethnic groupings included in the 2004 dataset created from the 2003 dataset, an alternative variable from 2003 was used to approximate ethnic groupings, allcult1. This variable asked which culture was dominant in each respondent’s personal life. When cross-checked against the ethnic groupings (White, Black, Asian and Other) recorded for each respondent in the 2003 dataset, all of the responses for allcult1 corresponded with the respondent’s ethnic grouping in these four categories.

The basic model uses four constraints: ethnicity, social class, age and sex. The main aim of SimHealth is to estimate the prevalence of type 2 diabetes at the output area level, using the 2003 and 2004 HSE and the 2001 Census data; these constraints are the most relevant risk factors for type 2 diabetes that are available at the individual level in both the Census and the HSE. Table CT003 from the 2001 Census provided the output area counts for the age, ethnic and sex categories. Table CS066 supplied the social grade classifications by both age and sex, although the age categories differed from the ones available for the ethnic groupings so were not included. Previous diabetes estimate models have included some measure of individual deprivation and created more accurate estimates than models with only age-sex-ethnicity distributions (Congdon, 2006). Both the aetiology literature pertaining to diabetes and this modelling evidence indicate that some measure of deprivation is needed; models which have not included it acknowledge that this data would increase the model goodness-of-fit (Forouhi et al., 2006).

The categorisation of constraints is important as well. Some of the constraints have very natural categories, such as sex (male or female), however, with other constraints there can be several viable combinations, such as ethnicity. Initially, the ethnic constraint was divided into two categories: white or non-white. However, these early simulations failed to reproduce diabetes prevalence well (based on the sample from Bradford). This was likely because the prevalence of type 2 diabetes among several non-white groups varies widely, with Asian groups having much higher rates than Chinese (Figure 3). Because SimHealth uses data from the HSE with disaggregated ethnicity variables, it was possible to further distinguish between White/Irish, Chinese/other, Black African, Black Caribbean, Indian, Pakistani, Bangladeshi.

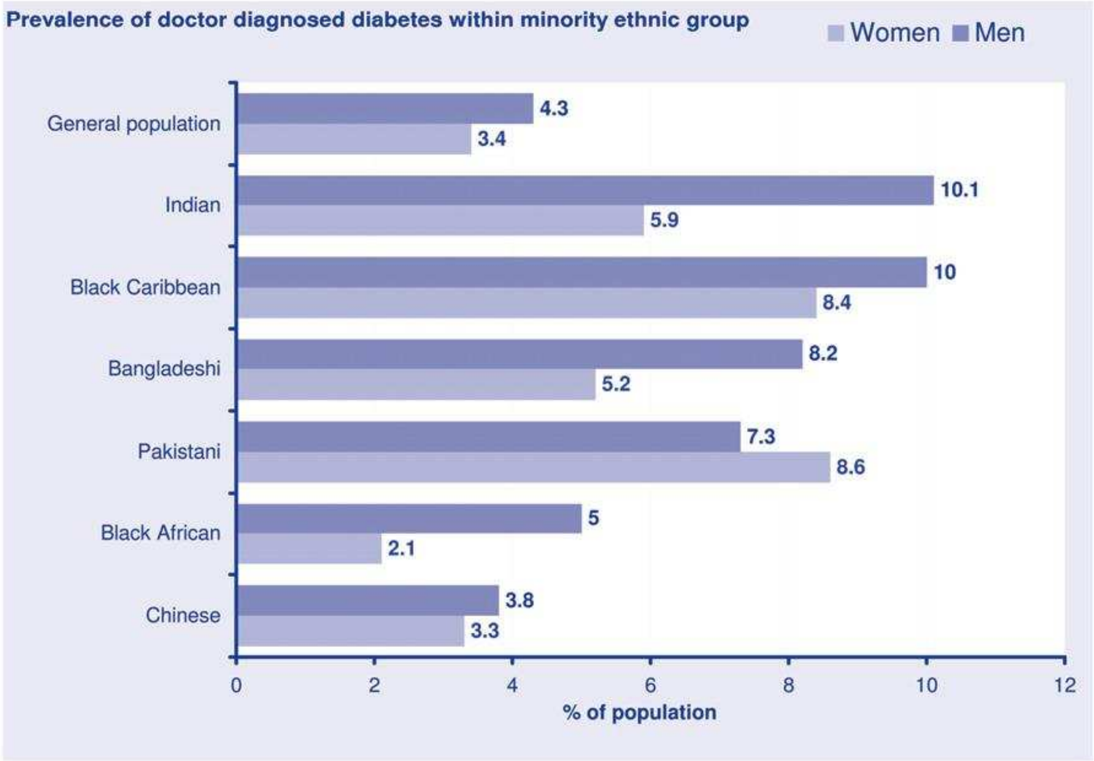

Figure 3. Diabetes prevalence by ethnicity. Reproduced from YPHO Diabetes Key Facts (Source: HSE 2004)

Within the ethnic categories, a number of decisions for categorisation were necessary. If a respondent was classified as white/Asian this was placed in the ‘other’ category. The ‘white’ category includes only people who have classified themselves as White British, White Irish, or Other White in the census or HSE. This is the best approximation for the purposes of our simulation, as non-whites do have a higher prevalence of diabetes. Any other ethnic combinations (e.g., Black/Indian) were also classified as ‘other’.

BMI is another category which proved difficult to combine. The HSE survey only calculates BMI for people aged 16 and over, although the 2004 dataset included BMI estimations for 2-15 year olds based on the UK standard of 1990 percentile curves (ERPHO, 2002). The SPSS syntax for this estimation was available; however, it used variables (day, month and age in years) that were not included in the public dataset so it was not possible to calculate the children’s BMIs from the 2003 dataset. There is debate over the appropriate age for BMI calculations with some experts arguing that adult BMI should only be calculated for people aged over 20 (see section 2.2.4 in the lit review); for the purpose of this analysis the cutoff of 16 that is used by HSE researchers will be maintained and children aged under 15 will be excluded from analysis for overweight and obesity.

The final 2003-4 dataset was created by selecting only those records with valid answers for the social grade, age, sex and ethnicity and people aged over 16 with valid responses for BMI; this adjusted the total population to 25,478 which includes the under-16s with unclassified BMIs. The remaining variables critical to the analysis and validation include tenure, marital status and diabetes type. The tenure and marital status variables are used in validation only; each of these are binary, with tenure coded as owned (outright or with a mortgage) or other (social or private rented, shared ownership, other) and marital status as either currently married or other (single, separated, divorced, widowed).

2.1 Population totals

The variables in the census need to be normalised by the true population as the differences in response rate for each variable lead to different total populations in each output area, depending on the variable in question. SimHealth uses the total population count from table CT003 as the base population for each output area, as some tables in the census have smaller totals due to non-response to a question. When the constraint tables are created for each variable, the total populations in each output area are normalised to match the total population in the output area as defined in CT003 (Figure 4). In Figure 4, cat 1-4 indicates the population in an OA that falls into one of four ethnic categories (white, black, Asian and other). Adj 1-4 represents the adjusted category populations, calculated by (cat1/sumpop)*realpop.

cat 1	**cat 2**	**cat 3**	**cat 4**	**sumpop**	**real pop**
314	3	3	7	327	325
adj1	**adj2**	**adj3**	**adj4**	**adjpop**	
312	2.9817	2.9817	6.9572	325	

Figure 4. Example of adjusted constraint variables

3 Validation

Each of the models discussed in this chapter were validated as described below. If the preliminary results from the sample dataset (2003 HSE dataset) did not meet the minimum validation criteria, the model was discarded and another configuration was tested.

3.1 Validation methods

Validation of microsimulation outputs is a vital aspect of the modelling procedure, however, very little literature includes any discussion of validation methods for synthetic population estimation (Voas & Williamson, 2000). The nature of microsimulation complicates the validation process, as the model outputs are estimates of unknown data. One commonly used approach to the validation is aggregation of the simulated data to a geographical level with known values for the constrained and unconstrained variables (Ballas & Clarke, 2001). To validate SimHealth, the individual-level output values are aggregated into OAs and the resulting percentages of population in categories of an unconstrained variable can be checked against the known values reported in the Census.

Each model created as part of this research is validated against both the constrained and unconstrained variables that are present in both the survey and the Census datasets, measuring the Total Absolute Error (TAE), Standardised Absolute Error (SAE) and percent error. The error between simulated populations from SimHealth and the actual census-defined populations is measured using TAE in the following equation (5):

$$\text{TAE} = \sum_{ij} \left| U_{ij} - T_{ij} \right| \qquad (5)$$

where U_{ij} is the observed count for the area i in category j

T_{ij} is the expected count for the area i in category j.

SAE is calculated as TAE divided by the total known (non-simulated) population for each area. In addition to TAE and SAE, percent error is often reported, which is SAE x 100. Voas and Williamson (2000) indicate that TAE and SAE are the most appropriate options for validating/evaluating estimated populations. In their discussion on evaluation of fit, the problem of validating microsimulation models is highlighted: "…no generally applicable methodology has emerged for measuring bias and variability." (Voas & Williamson, 2000 p.353).

The error thresholds for both stages need to be chosen based on the intended usage of the model. Because diabetes is a relatively rare disease (prevalence estimates are <4% of the general population), the model needs to be very accurate, with less than 10% error (SAE < 0.1) in 90% of the OAs for the constraints, and less than 20% error (SAE < 0.2) in 90% of the output areas for the unconstrained variables. If each of the tested models did not meet the criteria of less than 10% error in at least 90% of the areas for the constraint variables, then the model was discarded and another potential configuration was tested. These error thresholds are tighter than those usually used; often the models are expected to fit at least 80% of the areas with less than 20% error (Clarke and Madden, 2001). The final best-fitting model will create the population estimation from the full 2003 and 2004 HSE dataset.

Other options for error analysis include R^2, however, this method can inadvertently hide errors in some datasets; in SimHealth, TAE was quite high although R^2 appeared to be very good. The high TAE values can be masked if the simulated population is compared to the actual using a scatterplot and calculating R^2, as used by Ballas et al. (2005). The application of this error method was useful in

SimBritain (specifically, SimWales) because there were fewer and geographically larger areas; any large population loss would be easily identified in the scatterplots of model results. SimHealth uses much smaller and more numerous areas, so the magnitude of population loss would need to be much greater for it to be reflected in the R^2 (Figure 5)

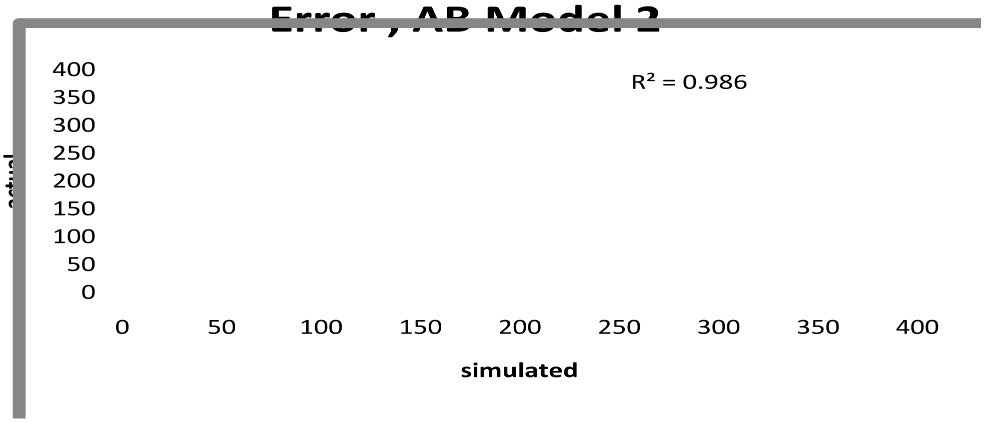

Figure 5. Error measured by R^2 (0.986)

The ability of R^2 error measures to mask error is clear when comparing Figures 5 and 6. The R^2 is very high, indicating that the model is a good fit. When the TAE is calculated for the same model outputs, it reflects the true amount of error in the simulations (Figure 6).

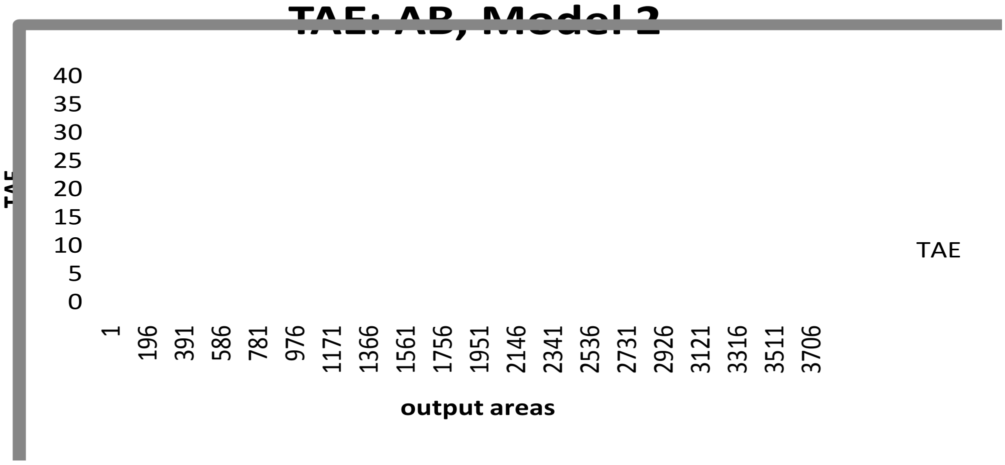

Figure 6. TAE (actual-simulated per OA)

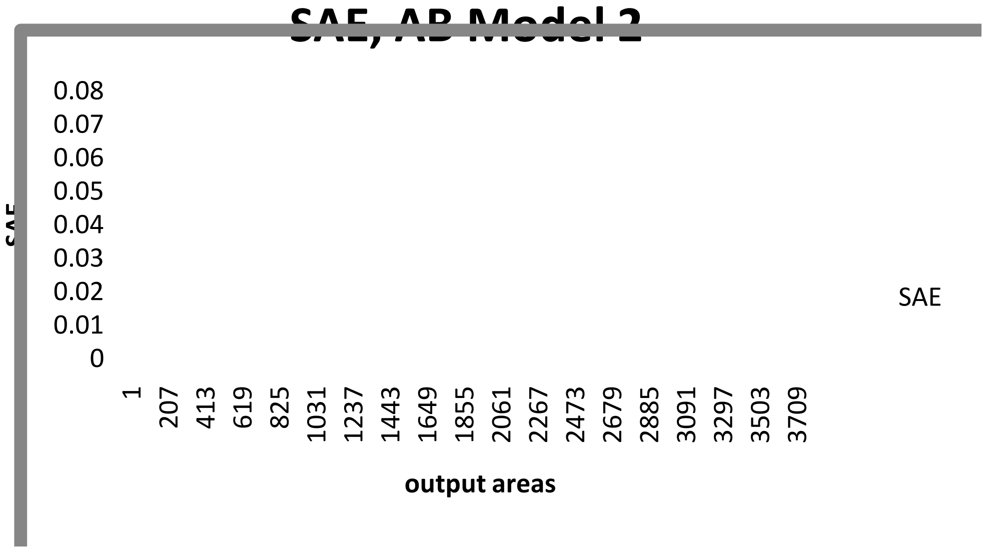

Figure 7. SAE for AB, Model 2, integerised

3.2 Validation variables

The validation of the model required the choice of variables with strong relationships to the constraints; if this was not the case, the validation would not be meaningful. Some modelling frameworks require that all of the variables be independent of each other, however, this is not feasible for this application (diabetes and obesity are themselves highly correlated). The validation process needs to include variables which are correlated to the health outcomes in order to assess the ability of the model to predict health outcomes; similarly, the constraint variables must also be correlated to the health outcomes.

To assess the relationships between the variables used in the population estimation, a series of correlation tests were carried out between each of the constraint and validation variables. A simple Chi-squared analysis with each of the variables coded into dichotomous categories showed that all of the variables were significantly associated with marital status and tenure at $p < .01$ with the exception of sex and tenure, which is significant at $p < .05$ (Table 3). Chi-square is a nonparametric statistical test designed for use with categorical variables that can identify whether the difference in distribution of data from more than one sample is due to chance or if they are significant. The magnitude of any identified significant relationships can be measured using Cramer's V (Field 2005).

Table 3. Chi-square statistics and significance (Cramer's V).

Variables (categories)	Marital Status	Tenure
Age (0-49, 50+)	3535.29 (.476)	158.05 (.101)
Sex (male, female)	17.00, (.033)	4.62 (.017)
Ethnicity (white, non-white)	24.02 (.039)	123.0 (.089)
Social Grade (A-C2, DE)	164.38 (.103)	1036.86 (.258)

Both tenure and marital status were initially selected as validation variables, however, tenure was difficult to model. The categories given for tenure in the HSE and Census did not match up; the only consistent categories between the two surveys were 'owned' (either outright or paying with a mortgage) and 'other' (all other categories). The following section will show that although the more aggregated constraints estimate the population well, further disaggregation is needed; the more precise the constraint categories are, the more precise the model can be when selecting suitable people to populate each area, leading to a more accurate synthetic population for representing the health outcomes. Marital status validated better against the simulated populations, probably because it was correlated to the variables constraining the population and is not spread relatively evenly throughout the population. Tenure had much higher error but there is less variation in tenure across the population; nearly 70% of the UK population own their home if the definition used in SimHealth's validation is applied, and over 75% of the HSE sample used in this analysis owned their own home (Census 2001).

4 Testing models

The initial model design matched the description of SimBritain: the model could be run for any number of geographical output units, with any number of survey respondents, and would reweight each individual iteratively against univariate constraints. The final 'new' weights for each person in each area, after the last constraint was reweighted, would then be sorted in ascending order and the decimal weights would be converted to integers; these final integer weights would add up to the total census-defined population in each geographical unit (Ballas et al., 2005). The variables which constrained the model (age, sex, ethnicity and social class) should reach a near-perfect fit with the reported 2001 Census population distributions.

Five conditions were examined with the intention of optimising SimHealth's population estimations: constraint categorisation, area clustering, integerisation of weights, initial weights, and creation of cross-tabulated constraint tables. Each of these conditions was adjusted in SimHealth, with the optimal choice recorded. With each subsequent adjustment, the SAE and percent error for the output areas were calculated and compared against previous SAE/percent error; this extensive, detailed testing led to the creation of the best configuration of SimHealth for the study area, and identified important contributions of this research to the field of synthetic population estimation. The following sections explain each adjustment and the impact of these changes on the final model design.

The variations on the baseline model specified for comparison include:

1. Experiment with different constraint configurations: Models 1-4
2. Run the configured models with clustered OAs
3. Remove the integerisation step
4. Adjust the initial weights to all equal 1 rather than the HSE-defined interview weight
5. Use a cross-tabulation routine to create probabilities for the interrelationships between the variables: feed these into the deterministic model to select out the correct individuals from the HSE

4.1 Configure constraints to fit specific population distributions

The order in which constraints are reweighted in SimHealth influence the accuracy of the final population estimation; the first constraint to be reweighted will be the most accurate. The study area includes many heterogeneous output areas (OAs), with some having older populations, high percentages of non-white residents or a mix of different social classes. This variety across the region makes it difficult to choose one order of constraint reweighting to most accurately estimate the population. To overcome this problem, four different constraint orders and categorisations were created, each to reflect different population characteristics (Table 4).

Table 4. Model variable combinations (number of categories)

	Model 1	Model 2/4	Model 3
Sex	Male, female (2)	Male, female (2)	Male, female (2)
Age	0-15, 16-49, 50+ (3)	0-15, 16-29, 30-49, 50-pensioner, pensioner+ (5)	0-15, 16-29, 30-49, 50-pensioner, pensioner+ (5)
Ethnicity	White, non-white (2)	White, Black, Asian, Other (4)	White, Black, Indian, Pakistani, Bangladeshi, Other (6)
Social Grade	A-C2, DE (2)	AB, C1, C2, D, E (5)	AB, C, DE (3)

Model 1 is the simplest configuration, with each constraint having only two or three categories. This configuration most closely matches the configurations used in SimBritain, as many of them are limited to two or three categories. Model 1 reweights individuals on the basis of ethnicity first, then social grade, age and finally sex. Ethnicity was listed first as an acknowledgement of its importance in predicting diabetes, which is the overall aim of SimHealth.

Model 2 is intended to be more precise with respect to age and to help best fit areas where there is low ethnic diversity and greater differences in social class (Table 8). Age, ethnicity and social grade are all disaggregated to create more accurate synthetic populations, as more detailed constraint categories will create a more accurate population. Ethnic distribution is disaggregated into white, black, Asian and other categories to provide a more detailed population profile than simply indicating the percent white and non-white, as in Model 1. Model 2 is identical in constraint configuration to Model 4, except Model 4 reweights individuals by age category first.

Model 3 is designed to best represent areas where there is greater ethnic diversity, and is important in accurately predicting type 2 diabetes, as different Asian ethnic groups have very different relative risks of type 2 diabetes (see Figure 3). Ethnicity is the most disaggregated variable in this model and is also the first constraint to be used in reweighting calculations. Social class is less detailed, although the age groupings are still divided into five groups. The following section on clustering indicates that output areas with the greatest ethnic diversity are also likely to be less diverse socially, so the simplified version of social grade is reasonable.

The initial models are compared on the basis of percent error: the minimum, maximum, mean, median and mode for each variable category in every model is compared. It is not possible to directly compare the constraint fit for each model because of the varied model configurations, however, the unconstrained variable categories were consistent across all models. The results from each of the model configurations are reported below (Tables 5-8). NB: in all of the following tables, (a) indicates that multiple modes existed with the reported one being the smallest.

Table 5. Model 1 validation statistics

Model 1	owned	other	unmarried	married
Mean	23.06	23.05	10.70	10.87
Median	18.57	18.40	9.09	9.18
Mode	7.58(a)	20.00(a)	.00	.00
Minimum	.02	.00	.00	.00
Maximum	78.40	78.98	48.74	49.10

Table 6. Model 2 validation statistics

Model 2	owned	other	unmarried	married
Mean	20.93	20.78	9.37	9.33
Median	20.21	19.65	7.88	7.95
Mode	25.00	.00 (a)	.00(a)	.00
Minimum	.00	.00	.00	.00
Maximum	65.39	66.83	39.64	37.55

Table 7. Model 3 validation statistics

Model 3	owned	other	unmarried	married
Mean	23.20	22.56	10.54	11.54
Median	18.50	17.71	8.59	10.20
Mode	5.26(a)	18.18	.00	.00
Minimum	.00	.00	.00	.00
Maximum	86.11	87.22	46.18	46.44

Table 8. Model 4 validation statistics

Model 4	owned	other	unmarried	married
Mean	22.82	22.17	7.90	8.29
Median	18.28	18.06	6.48	7.03
Mode	7.14(a)	14.08(a)	.00	.00
Minimum	.00	.06	.00	.00
Maximum	79.38	84.66	47.84	46.39

If only the unconstrained variables are compared, model 4 seems to provide the best fit, then models 2, 1 and 3 in that order. However, these models may still have significant errors in some of the output areas, as shown by the very high maximum percent errors in each model configuration.

After this round of simulation experiments, the problems inherent in proportional fitting (assumptions that all areas have similar populations) proved too difficult to overcome using only constraint re-categorisation. The decision to identify and group similar areas to simulate concurrently resulted from literature on cluster analysis; the next model improvement was designed to decrease error through clustering of the output areas, and selecting the optimal model configuration (Model 1-4) to simulate the population in each cluster.

4.2 Improving model fit using clustering

A problem in using proportional fitting is that the calculations begin with the assumption that all areas adjusted to fit some pre-defined row and column totals have the same initial value (Table 9) (Norman, 1999). With an iterative process, or with many areas which have similar characteristics, this assumption may not affect the analysis. Unfortunately, the number and diversity of areas included in this study meant that the model attempted to ‘smooth’ the population distributions of each constraint towards a global mean. This is not a challenge that has been acknowledged in previous reweighting research, however, one solution is to cluster the output areas to create aggregate groups with shared characteristics.

Table 9. The initial weight assumptions for IPF

	White	Non-white	Row totals
Male	1.0	1.0	**0.5**
Female	1.0	1.0	**0.5**
Column totals	**0.75**	**0.25**	**1.00**

All of the cluster analysis was carried out in SPSS 13.0 using 2001 census output area data, to be consistent with SimHealth. First, a two step cluster analysis was carried out to identify the optimal number of clusters from the dataset. Because the two step analysis is not the most appropriate methods for creating clusters from large datasets, k-means cluster analysis was carried out to identify the final clusters. SPSS allows for two methods of k-means cluster analysis: iterative, where the centres are updated with each iteration, or classification only which does not give you information on the cluster centres (preventing a direct comparison of the cluster attributes). The use of k-means cluster analysis to identify groupings of most similar output areas is an established methodology in geodemograhpic research; Vickers et al. (2005) use this clustering technique in the national output area classification for England. The variables used as constraints were chosen to also constrain the clusters with a maximum of five clusters allowed. The clusters are based on the percent of the population in each of four categories: social grade D or E, over 50 years of age, non-white and male. These criteria are based on the risk factors for diabetes, although there is little variation between male and female risk.

Using a 2 step cluster analysis, 5 clusters were identified as being the natural clusters. An iterative k-means cluster analysis was then used to create five clusters. The cluster membership varied from 302 records to 1,304. The table below shows the distances between the final cluster centres (classification centres) (Table 10). Clusters 3 and 4 are the most dissimilar and clusters 2 and 3 are the most similar.

Table 10. Distances between final cluster centres

Cluster	1	2	3	4	5
1		46.39	33.83	69.13	29.08
2	46.39		18.64	66.04	27.29
3	33.83	18.64		70.64	28.56
4	69.13	66.04	70.64		56.93
5	29.08	27.29	28.56	56.93	

Once the clusters were identified the main characteristics of each was clear (Table 11). The cells highlighted in yellow indicate the cluster with the greatest percent of the total population in that category, with the lowest in blue.

Table 11. Cluster characteristics

Cluster	% DE	% nonwhite	% over50	% male
1	62.23	5.39	52.34	44.42
2	22.72	5.76	28.51	49.12
3	29.31	3.02	45.65	47.52
4	51.04	64.61	18.79	48.92
5	49.79	8.22	26.45	47.88
all	**37.76**	**11.07**	**32.26**	**48.05**

Overall, the clusters can be defined as follows:

Cluster1: High percent in social grade DE and over the age of 50, low ethnic diversity
Cluster 2: Low percent in DE, over 50, ethnic diversity
Cluster3: Low percent in DE, ethnic diversity, higher percent over 50
Cluster 4: Highest ethnic diversity and percent DE, young population
Cluster 5: High percent DE, average ethnic diversity, low % over 50

Dominant cluster characteristics include:

Cluster 1: Aged, deprived
Cluster 2: Affluent
Cluster 3: Low ethnic diversity
Cluster 4: Young, ethnically diverse, more deprived
Cluster 5: similar to 4 but less ethnically diverse

The clustering analysis allowed for the identification of the optimal model configuration for areas with similar attributes. This was achieved by comparing the mean, minimum, maximum percent error along with standard error of the mean for each model. A comparison of the unconstrained variable, marital status, across all of the clusters and model configurations showed that some models were more accurate for each of the clusters. Model 1, when validated using marital status, had high error for all of the clusters. As expected, the most ethnically diverse area (Cluster 4) was best modelled using Model 3, which has the most disaggregated ethnic categories. Overall the ‘best’ model configuration for each cluster was the one with lowest average percent error that had a median very close to the mean, a low value for the mode and low minimum and maximum percent error. The minimum and maximum values were the least important with the most importance placed on mean and median percent error. The variation in outputs (percent error for unmarried) for each model configuration is shown in Tables 12-16.

Table 12. The cluster 1 population was best modelled using configuration 2

Unmarried % error	Model1	Model2	Model3	Model4
Mean	13.64	8.78	13.87	17.14
Std. Error of Mean	.47	.40	.46	.58
Median	13.18	7.20	13.14	15.85
Mode	.00(a)	.06(a)	.03(a)	.15(a)
Minimum	.00	.06	.03	.15
Maximum	41.55	32.54	39.12	47.30

Table 13. Cluster 2 population was best modelled using configuration 4

Unmarried % error	Model1	Model2	Model3	Model4
Mean	7.87	8.06	7.81	6.75
Std. Error of Mean	.19	.17	.20	.13
Median	6.22	6.80	6.11	6.15
Mode	8.33	.00(a)	.00	.00(a)
Minimum	.00	.00	.00	.00
Maximum	42.30	36.56	43.41	25.74

Table 14. Cluster 3 population was best modelled using configuration 4

Unmarried % error	Model1	Model2	Model3	Model4
Mean	8.32	10.78	8.30	6.70
Std. Error of Mean	.21	.23	.20	.19
Median	7.70	10.56	7.44	5.82
Mode	1.06	13.97(a)	1.28(a)	.00(a)
Minimum	.00	.00	.08	.00
Maximum	30.10	28.41	31.41	47.84

Table 15. The population of cluster 4 was best modelled using configuration 3

Unmarried % error	Model1	Model2	Model3	Model4
Mean	8.26	6.96	6.98	10.52
Std. Error of Mean	.32	.36	.30	.36
Median	7.59	4.97	6.23	9.98
Mode	.00(a)	.06(a)	.00(a)	.02(a)
Minimum	.00	.06	.00	.02
Maximum	32.28	33.27	34.35	30.27

Table 16. The population of cluster 5 was best modelled using configuration 4

Unmarried % error	Model1	Model2	Model3	Model4
Mean	15.31	10.75	15.20	6.85
Std. Error of Mean	.27	.24	.27	.17
Median	14.38	9.10	14.48	5.51
Mode	13.65	11.30	.05(a)	.00
Minimum	.05	.00	.05	.00
Maximum	48.74	39.64	46.18	41.33

The validated populations, using the best fitting model configurations for each cluster, were a great improvement over the early model runs with all of the output areas reweighted together. The next element of the modelling process to come under scrutiny was the integerisation of the initial decimal weights whole numbers. Although integerisation was a logical step in SimBritain, this process may prove unnecessary for SimHealth.

4.3 Intergerisation

The integerisation process described by Ballas et al. (2005) was selected for use in SimBritain following extensive testing. This method is intended to be carried out for each geographical unit in turn (Ballas et al., 2005 p.40):

- Create two variables named counter and cumulative weight and set them to zero
- Sort all individuals into ascending order of the new weights
- Increase the cumulative weight by the weight of the next individual
- If cumulative weight > 1, set the counter to an integer weight equal to the rounded weight value and subtract this value from cumulative weight. Increase counter by 1 and move to the next individual.
- If counter<total individuals, return to step 3, else quit.

SimBritain was created to reweight entire households from a very small sample of the population (using the British Household Panel Survey (BHPS)) to Census wards with large populations (approximately 13,000 individuals); SimHealth uses a large population sample (n=30,297) from which around 250-300 individuals are selected to populate each Census output area. When the reweighting process is carried out, the resulting ratio of individuals from the HSE who represent the constraint category from the Census is very small. When the method of integerisation as suggested by Ballas et al. (2005) was implemented, many of the very small weights were discarded and the resulting populations had very high TAE and SAE (Figures 8 and 9).

The integerisation process was removed from the code, with positive results; the constraint variables all had error rates below 10 (SAE<0.10). Figures 8-9 compare the percent error (SAE x 100) for integerised and non-integerised outputs using the Model 2 configuration.

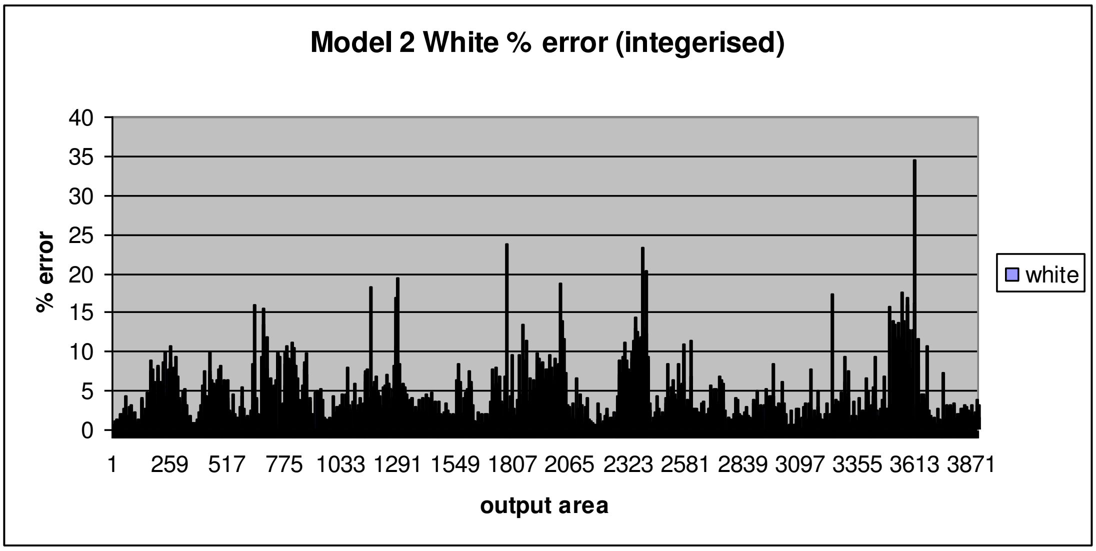

Figure 8. Percent error from integerised outputs

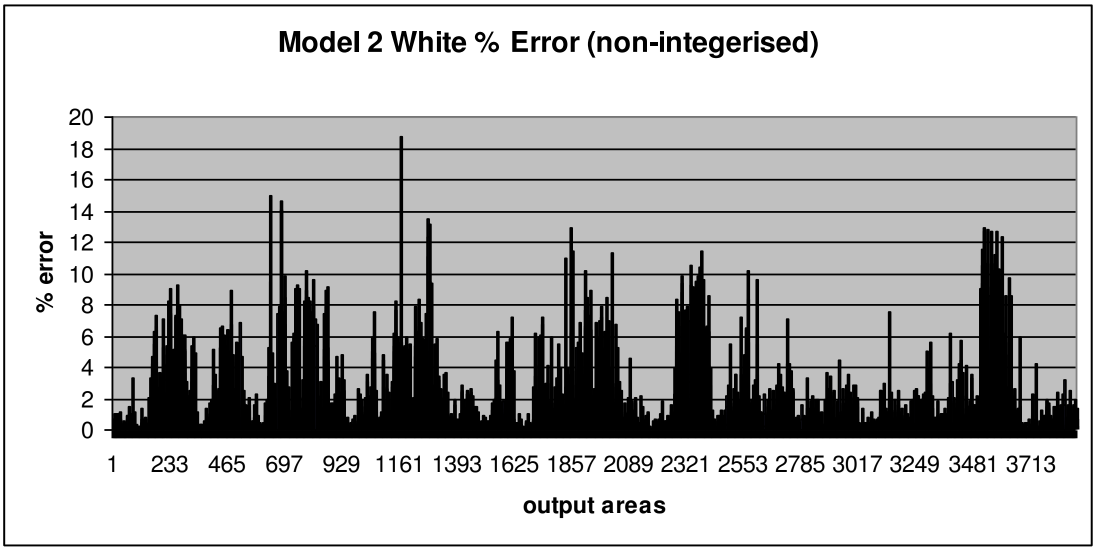

Figure 9. Percent error from non-integerised outputs

Following detailed comparisons of the clusters with the non-integerised and integerised versions of SimHealth, it was found that removing the integersiation step did improve population estimates in some of the clusters Tables 17-.21). The population estimates were the same for model configurations 2 and 4, and the paired variables (married and unmarried, owned and other) also had the same percent error values for each cluster-model grouping.

Table 17. Cluster 1 unintegersied model comparison.

Cluster 1 unintegersied	owned1	marry1	owned2	marry2	own3	marry3	owned4	marry4
Mean	32.37	12.76	26.22	8.06	30.36	10.16	26.22	8.06
Std. Error of Mean	.962	.50	.87	.40	.93	.45	.87	.40
Median	32.44	11.63	26.06	6.68	29.95	8.99	26.06	6.68
Mode	.77(a)	.06(a)	.09(a)	.07(a)	.17(a)	.00(a)	.09(a)	.07(a)
Minimum	.77	.06	.09	.07	.17	.00	.09	.07
Maximum	66.70	47.40	59.57	40.86	64.17	43.40	59.57	40.86

Table 18. Cluster 2 unintegersied model comparison.

Cluster 2 unintegersied	owned1	marry1	owned2	marry2	own3	marry3	owned4	marry4
Mean	16.30	8.20	17.82	7.64	16.55	6.73	17.82	7.64
Std. Error of Mean	.24	.18	.23	.12	.23	.12	.23	.12
Median	16.95	6.94	18.83	7.48	17.15	6.30	18.83	7.48
Mode	.03(a)	.00(a)	.00(a)	.01(a)	.13(a)	.00(a)	.00(a)	.01(a)
Minimum	.03	.00	.00	.01	.13	.00	.00	.01
Maximum	67.30	38.83	54.49	26.67	57.81	28.77	54.49	26.67

Table 19. Cluster 3 unintegersied model comparison.

Cluster 3 unintegersied	owned1	marry1	owned2	marry2	own3	marry3	owned4	marry4
Mean	15.79	7.28	18.17	8.96	16.59	7.20	18.17	8.96
Std. Error of Mean	.27	.18	.30	.18	.29	.16	.30	.18
Median	16.99	6.86	19.79	9.20	17.94	7.03	19.79	9.120
Mode	.20(a)	.05(a)	.01(a)	.00(a)	.11(a)	.03(a)	.01(a)	.00(a)
Minimum	.20	.05	.01	.00	.11	.03	.01	.00
Maximum	50.40	42.52	43.61	22.39	46.63	34.57	43.61	22.39

Table 20. Cluster 4 unintegersied model comparison.

Cluster 4 unintegerised	owned1	marry1	owned2	marry2	own3	marry3	owned4	marry4
Mean	20.35	9.51	14.85	7.17	14.77	7.60	14.85	7.17
Std. Error of Mean	.65	.32	.57	.22	.64	.24	.57	.22
Median	20.00	9.30	13.21	7.13	12.23	7.81	13.21	7.13
Mode	.11(a)	.08(a)	.03(a)	.08(a)	.01(a)	.12(a)	.03(a)	.08(a)
Minimum	.11	.08	.03	.08	.01	.12	.03	.08
Maximum	55.32	32.99	51.31	17.66	54.85	18.47	51.31	17.66

Table 21. Cluster 5 unintegersied model comparison.

Cluster 5 unintegerised	owned1	marry1	owned2	marry2	own3	marry3	owned4	marry4
Mean	25.11	8.46	21.67	5.32	23.11	5.60	21.67	5.32
Std. Error of Mean	.49	.25	.42	.13	.45	.14	.42	.13
Median	21.69	5.95	19.37	4.29	20.29	4.40	19.37	4.29
Mode	.03(a)	.01(a)	.11(a)	.00(a)	.02(a)	.00(a)	.11(a)	.00(a)
Minimum	.03	.01	.11	.00	.02	.00	.11	.00
Maximum	66.72	41.27	60.46	28.58	63.65	30.34	60.46	28.58

Table 22 shows a direct comparison between the integerised and non-integerised model adjustment. Every cluster is listed with the optimal model configuration and the percent error for marital status including mean, median, mode, minimum and maximum (Table 22).

Table 22. Percent error: mean, median, mode, minimum, maximum

Integerised			Non-integerised	
	Optimal model	Percent error characteristics	Optimal model	Percent error characteristics
Cluster 1	2	8.78; 7.19; 0.6; 0.6; 32.54	2	8.06; 6.68; .07; .07; 40.86
Cluster 2	4	6.75; 6.15; 0; 0; 25.74	3	6.73; 6.3; 0; 0; 28.77
Cluster 3	4	6.70; 5.82; 0; 0; 47.84	3	7.2; 7.03; 0; 0; 34.57
Cluster 4	3	6.97; 6.23; 0; 0; 34.35	4	7.17; 7.13, .08; .08; 17.66
Cluster 5	4	6.84; 5.51; 0; 0; 41.33	4	5.32; 4.29; 0; 0; 28.58

The version of SimHealth with specified constraint configurations and non-integerised final weights provided an improved model fit for some clusters when evaluated using TAE, SAE and percent error. Two additional adjustments still need to be evaluated: initial weights and cross-tabulation.

4.4 HSE-defined weights

A potential difficulty with the deterministic reweighting methodology is the use of initial weights (which represent a probability of each person responding to the initial survey) in the calculation (indicated by w_i in equation 1); is the use of a non-response weight suitable in spatial microsimulation, since the process of deterministic reweighting inherently chooses and reweights individuals to be representative of the small area? To answer this question, it is important to understand the methods used to select respondents and subsequently create weights for the national survey.

There are certain populations over- and under-represented in each year of the HSE (table 23). The selection method used in collecting responses can begin to explain the inability of unweighted data to accurately depict the population; survey interviewers may visit homes during typical working hours, missing out individuals who work outside of the home. Another potential reason is simply that people in certain age-sex groups are less willing to complete the survey.

Table 23. Misrepresentation in the HSE. (Source: DOH 2003,2004)

HSE YEAR	Over-represented	Under-represented
2003	Men aged over 55 Women aged under 25	Men aged under 35 Women aged over 55
2004	Men and women aged over 55	Men and women aged under 25

The HSE includes four weights with the interview weight suggested as the appropriate weight to use in analysis of individual-level data. The interview weight was introduced in the 2003 survey and continued in the 2004 survey to correct for non-response bias; the population of respondents is also adjusted, using the household weight, to fit mid-year estimates of age-sex distributions at the level of Government Office Region (GOR) (DOH 2003, 2004). The probability of response, which was needed to generate the weights in both datasets, was calculated using a logistic regression model (response/non-response) that included age-sex interaction, age group, sex, GOR, household type, and the social class of the household representative person (HRP) (DOH, 2003). For both years a maximum of 3 households for each address were selected for interview, with weights calculated for each household to ensure accurate representation of the age-sex distribution in each GOR, however, all adults and children in each household are given identical weights: these are the final household weights. The 2003 weights, both household and interview, were ‘trimmed’ to remove the weights below the first and above the 99th percentile with the intention of removing outliers (DOH, 2003). (DOH, 2003, 2004).

In both 2003 and 2004, the resulting re-weighted population distribution for the age-sex groupings matched the known population distribution well in all GORs. This is important for researchers who are using the datasets to compare across areas, or who intend to compare new data against previous years. The interview weights for 2003 ranged from 0.39 to 3.2, which is reasonable for use in the model. The 2004 general population weights range from 6.07-50.67 and the ethnic boost sample ranges from 0.08 to 20.24, which are much higher than would normally be seen in a microsimulation model for the starting weight. The combined 2003-4 population (with respondents excluded if they did answer constraining questions or the BMI measure) includes 25, 457 individuals with weights ranging from 0.08 to 39.16.

An alternative option is to set all of the initial weights to 1, creating a uniform starting point for the reweighting process. Because the deterministic reweighting procedure only chooses individuals that are representative of an area (based on the constraint distribution for the small area), there should be no need to use the HSE weights. Returning to the previous studies completed by Ballas et al, there is not a clear reason for the use of survey-produced weights. The earlier studies have all reweighted the BHPS, which utilises a different method of weight calculation with the same intention: to correct for non-response (both households and individuals within each household) and account for sampling design (Ballas et al., 2005). Adjusting all of the initial weights to 1 was a simple change. The results were positive, with an improvement on the validation over the previous models for clusters 1, 3 and 4 (Table 24). This may underestimate the effect on the pooled 2003-4 HSE dataset, however, as the range of weight values is much greater for 2004, due to the weighting procedure. These weights may not be appropriate for the purposes used here, as they could over-bias certain people with the highest weights, skewing the resulting simulated population.

Table 24. Percent error: mean, median, mode, minimum, maximum

Non-integerised (weight of 1)			**Non-integerised (HSE weight)**	
	Optimal model	**Percent error characteristics**	**Optimal model**	**Percent error characteristics**
Cluster 1	2	7.18; 6.15; .02; .02; 41.20	2	8.06; 6.68; .07; .07; 40.86
Cluster 2	4	7.35; 6.96; 0; 0; 29.16	3	6.73; 6.3; 0; 0; 28.77
Cluster 3	1	6.84; 6.25; .02; .02; 41.45	3	7.2; 7.03; 0; 0; 34.57
Cluster 4	3	5.93; 5.8; .02; .02; 20.16	4	7.17; 7.13, .08; .08; 17.66
Cluster 5	4	5.59; 4.41; 0; 0; 30.41	4	5.32; 4.29; 0; 0; 28.58

A comparison between each of the available models for each cluster is shown in tables 25-29. Clusters 1 and 5 were clearly simulated most accurately in one of the model/configuration combinations, however, the distinction between model goodness-of-fit was not as obvious for the remaining three clusters. In these cases the standard deviation of the percent error from the mean was used as a deciding factor.

Table 25. Cluster 1 model comparison

integerised	Model1	Model2	Model3	Model4
Mean	13.64	8.78	13.87	17.14
Median	13.18	7.20	13.14	15.85
Mode	0.00	0.06	0.03	0.15
Minimum	0.00	0.06	0.03	0.15
Maximum	41.55	32.54	39.12	47.30
Std. Deviation	8.09	6.74	7.87	9.12
nonint HSE wt				
Mean	12.76	8.06	10.16	8.06
Median	11.63	6.68	8.99	6.68
Mode	0.06	0.07	0.00	0.07
Minimum	0.06	0.07	0.00	0.07
Maximum	47.40	40.86	43.40	40.86
Std. Deviation	8.62	6.94	7.75	6.94
nonint wt 1				
Mean	13.17	7.18	13.97	8.13
Median	12.03	6.15	12.67	6.89
Mode	0.11	0.02	0.08	0.05
Minimum	0.11	0.02	0.08	0.05
Maximum	48.11	41.20	49.35	41.62
Std. Deviation	8.61	5.77	8.89	7.00

Table 26. Cluster 2 model comparison

Integerised	Model1	Model2	Model3	Model4
Mean	7.87	8.06	7.81	6.75
Median	6.21	6.80	6.11	6.15
Mode	8.33	0.00	0.00	0.00
Minimum	0.00	0.00	0.00	0.00
Maximum	42.30	36.56	43.41	25.74
Std. Deviation	6.99	6.29	7.11	4.52
nonint HSE wt				
Mean	8.20	7.64	6.73	7.64
Median	6.94	7.48	6.30	7.48
Mode	0.00	0.01	0.00	0.01
Minimum	0.00	0.01	0.00	0.01
Maximum	38.83	26.67	28.77	26.67
Std. Deviation	6.48	4.48	4.17	4.48
Nonint wt 1				
Mean	7.95	7.25	7.54	7.35
Median	6.57	6.50	7.17	6.96
Mode	0.00	0.01	0.02	0.00
Minimum	0.00	0.01	0.02	0.00
Maximum	40.72	29.16	30.08	29.16
Std. Deviation	6.82	4.90	4.58	4.46

Table 27. Cluster 3 model comparison

Integerised	Model1	Model2	Model3	Model4
Mean	8.32	10.78	8.30	6.70
Median	7.69	10.56	7.44	5.82
Mode	1.06	13.97	1.28	0.00
Minimum	0.00	0.00	0.08	0.00
Maximum	30.10	28.41	31.41	47.84
Std. Deviation	5.68	6.39	5.45	5.14
Nonint HSE wt				
Mean	7.28	8.96	7.20	8.96
Median	6.86	9.20	7.03	9.20
Mode	0.05	0.00	0.03	0.00
Minimum	0.05	0.00	0.03	0.00
Maximum	42.52	22.39	34.57	22.39
Std. Deviation	5.03	5.05	4.54	5.05
Nonint wt 1				
Mean	6.84	8.03	7.85	8.58
Median	6.25	7.31	7.63	8.66
Mode	0.02	0.01	0.01	0.01
Minimum	0.02	0.01	0.01	0.01
Maximum	41.45	41.62	32.26	21.74
Std. Deviation	4.95	5.36	4.74	4.95

Table 28. Cluster 4 model comparison

Integerised	Model1	Model2	Model3	Model4
Mean	8.26	6.96	6.97	10.52
Median	7.59	4.97	6.23	9.98
Mode	0.00	0.06	0.00	0.02
Minimum	0.00	0.06	0.00	0.02
Maximum	32.28	33.27	34.35	30.27
Std. Deviation	5.92	6.55	5.61	6.68
Nonint HSE wt				
Mean	20.35	7.17	7.60	7.17
Median	20.00	7.13	7.81	7.13
Mode	0.11	0.08	0.12	0.08
Minimum	0.11	0.08	0.12	0.08
Maximum	55.32	17.66	18.47	17.66
Std. Deviation	12.01	4.14	4.46	4.14
Nonint wt 1				
Mean	9.16	6.87	5.93	6.12
Median	8.61	6.69	5.80	5.92
Mode	0.06	0.04	0.02	0.02
Minimum	0.06	0.04	0.02	0.02
Maximum	34.95	21.74	20.16	15.70
Std. Deviation	6.06	4.03	3.78	3.72

Table 29. Cluster 5 model comparison

Integerised	**Model1**	**Model2**	**Model3**	**Model4**
Mean	15.31	10.75	15.20	6.84
Median	14.38	9.10	14.48	5.51
Mode	13.65	11.30	0.05	0.00
Minimum	0.05	0.00	0.05	0.00
Maximum	48.74	39.64	46.18	41.33
Std. Deviation	9.31	8.20	9.13	5.78
Nonint HSE wt				
Mean	8.46	5.32	5.59	5.32
Median	5.95	4.29	4.40	4.29
Mode	0.01	0.00	0.00	0.00
Minimum	0.01	0.00	0.00	0.00
Maximum	41.27	28.58	30.34	28.58
Std. Deviation	8.54	4.45	4.81	4.45
Nonint wt 1				
Mean	8.99	6.09	6.88	5.59
Median	6.47	5.02	5.68	4.41
Mode	0.01	0.00	0.01	0.00
Minimum	0.01	0.00	0.01	0.00
Maximum	42.80	29.09	35.06	30.41
Std. Deviation	8.93	4.66	5.56	4.78

From the comparisons of available univariate models, each of the clusters does have one model configuration and weighting/integerisation scheme which results in the lowest error for the validation. All of the clusters were best simulated using nonintegerised models, so the cross-tabulated run will also use nonintegersied methods. Clusters 1and 4 populations were simulated most accurately when an initial weight of 1 was used, rather than the HSE initial weight value. Clusters 2, 3 and 5 had a better fit with the HSE initial weight, but the potential error caused by the high weight values in the 2004 dataset needs to be considered further. Model configuration 2 gave the lowest error for clusters 1 and 5 and configuration 3 was best suited to clusters 2, 3 and 4. The final adjustment to the models, cross-tabulation, will be tested for each cluster using the optimal model configuration.

4.5 Using cross-tabulation to improve fit

The biggest challenge with the current model is the lack of cross-tabulated data at the individual level, which would provide better detail about the most suitable individuals from the HSE; if we know that an area has a total population of 200, of which 150 are white, 90 are male, and 50 are over the age of 25, we still do not know how many of those people are white males over the age of 25. We can offer a guess, but with the current method of deterministic reweighting, each person from the survey who falls into any of the categories (white, male, over 25) has an equal chance of being selected.

For example, if the first constraint is ethnicity, then let’s assume the model predicts with 100% accuracy and assigns probabilities that add up to 150 white individuals and 50 non-white people. This will continue with the other constraint variables, with each person given a decimal weight that is later converted to an integer. The result is that after integerisation of weights often the first constraint variable will match up with great accuracy, but subsequent constraint variables will have a greater level of error. This greater error occurs because each constraint variable assigns weights to individuals in isolation, so there is no way to account for any inter-relationship (such as the age, sex and ethnicity of each person in an area) even if it is known.

The census tables used to create the constraint tables were not univariate; they included two-dimensional cross tabulations between constraint variables such as ethnicity by sex and ethnicity by age categories (tables CT003 and CS066 from the 2001 Census). The reweighting method used initially for SimHealth does not allow for any relationships between the variables; each variable is reweighted to fit each output area in isolation. This loss of known inter-variable relationships causes the model to be less accurate than possible. A method was devised to create joint probability distributions between all four of the constraint variables.

IPF is frequently used to reweight the probabilities of an even occurring in a smaller geographical area based on known distributions. Some inter-constraint relationships were already known from the census tables so the method was simplified from IPF to multiplication of probability distributions with the aim of creating new distributions (equations 2, 3):

$$\mathbf{P(A)\ x\ P(E,S) = P(A,E,S)} \qquad (2)$$

$$\mathbf{P(A,E,S)\ x\ P(SG) = P(SG,A,E,S)} \qquad (3)$$

where **P(x)** is the probability of **x** occurring within a given area and
A is the age category
E is the ethnic category
S represents the sex of each individual and
SG is the social grade of each respondent.

Using the Census datasets (CT003 and CS066), the probability of any age or ethnic by sex classification was calculated by dividing the number of people in each category (e.g., for sex this would be the number of males or females in the output area) by the total population of the area. In all cases, the probabilities of P(E,S) and P(A) were then multiplied together to calculate the probability of an individual fitting into any P(A,E,S) distribution. Once P(A,E,S) was known for each output area, it was multiplied by P(SG) to reach the final distribution of P(SG,A,E,S) (see Tables 30-1 for an example).

Table 30. Cross-tabulation example

probabilities	male	female	white	nonwhite
original	0.55	0.45	0.7	0.3

Can be cross-tabulated to produce:

Table 31. Cross-tabulation result

	male	**female**
White	0.55 x 0.7	0.45 x 0.7
nonwhite	0.55 x 0.3	0.45 x 0.3

The only adjustment made to the census tables was the standardisation of each constraint category to the population of the output area as defined in CT003001 ("all people"). Table CS066 only included individuals age 16 and over, so the populations from this table were standardised by the total population from table CT003, reflecting the proportion of the total population included in each social grade. The final cross-tabulated constraint tables for Models 1, 2 and 3 included 24, 200 and 180 categories respectively.

The test run of the cross-tabulated data showed that the cross-tabulation did not improve the ability of SimHealth to accurately predict population characteristics. For each of the clusters, both the optimal model and Model 1 configurations were run, and the results for the unmarried validation category were compared. A comparison of the model outputs shows that the univariate models provided a better fit with marital status than the cross-tabulated models (Tables 31-35).

Table 32. Complete model comparisons, Cluster 1

integerised	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	13.64	8.78	13.87	17.14
Median	13.18	7.20	13.14	15.85
Mode	0.00	0.06	0.03	0.15
Minimum	0.00	0.06	0.03	0.15
Maximum	41.55	32.54	39.12	47.30
Std. Deviation	8.09	6.74	7.87	9.12
nonint HSE wt	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	12.76	8.06	10.16	8.06
Median	11.63	6.68	8.99	6.68
Mode	0.06	0.07	0.00	0.07
Minimum	0.06	0.07	0.00	0.07
Maximum	47.40	40.86	43.40	40.86
Std. Deviation	8.62	6.94	7.75	6.94
nonint wt 1	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	13.17	7.18	13.97	8.13
Median	12.03	6.15	12.67	6.89
Mode	0.11	0.02	0.08	0.05
Minimum	0.11	0.02	0.08	0.05
Maximum	48.11	41.20	49.35	41.62
Std. Deviation	8.61	5.77	8.89	7.00
Crosstab nonint	**Model1HSE**	**Model1wt1**	**Model2wt1**	**Model2HSE**
Mean	14.27	13.43	13.17	15.40
Median	13.20	12.27	12.53	14.99
Mode	0.03	0.06	0.09	0.01
Minimum	0.03	0.06	0.09	0.01
Maximum	47.64	48.00	40.73	42.83
Std. Deviation	8.99	8.66	7.60	8.06

Table 33. Complete model comparisons, Cluster 2

Integerised	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	7.87	8.06	7.81	6.75
Median	6.21	6.80	6.11	6.15
Mode	8.33	0.00	0.00	0.00
Minimum	0.00	0.00	0.00	0.00
Maximum	42.30	36.56	43.41	25.74
Std. Deviation	6.99	6.29	7.11	4.52
nonint HSE wt	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	8.20	7.64	6.73	7.64
Median	6.94	7.48	6.30	7.48
Mode	0.00	0.01	0.00	0.01
Minimum	0.00	0.01	0.00	0.01
Maximum	38.83	26.67	28.77	26.67
Std. Deviation	6.48	4.48	4.17	4.48
Nonint wt 1	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	7.95	7.25	7.54	7.35
Median	6.57	6.50	7.17	6.96
Mode	0.00	0.01	0.02	0.00
Minimum	0.00	0.01	0.02	0.00
Maximum	40.72	29.16	30.08	29.16
Std. Deviation	6.82	4.90	4.58	4.46
Crosstab nonint	**Model1HSE**	**Model1wt1**	**Model3wt1**	**Model3HSE**
Mean	8.22	7.98	7.54	7.29
Median	7.18	6.58	5.89	6.24
Mode	0.02	0.00	0.02	0.02
Minimum	0.02	0.00	0.02	0.02
Maximum	33.53	40.46	37.89	27.46
Std. Deviation	5.98	6.81	6.49	5.41

Table 34. Complete model comparisons, Cluster 3

Integerised	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	8.32	10.78	8.30	6.70
Median	7.69	10.56	7.44	5.82
Mode	1.06	13.97	1.28	0.00
Minimum	0.00	0.00	0.08	0.00
Maximum	30.10	28.41	31.41	47.84
Std. Deviation	5.68	6.39	5.45	5.14
Nonint HSE wt	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	7.28	8.96	7.20	8.96
Median	6.86	9.20	7.03	9.20
Mode	0.05	0.00	0.03	0.00
Minimum	0.05	0.00	0.03	0.00
Maximum	42.52	22.39	34.57	22.39
Std. Deviation	5.03	5.05	4.54	5.05
Nonint wt 1	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	6.84	8.03	7.85	8.58
Median	6.25	7.31	7.63	8.66
Mode	0.02	0.01	0.01	0.01
Minimum	0.02	0.01	0.01	0.01
Maximum	41.45	41.62	32.26	21.74
Std. Deviation	4.95	5.36	4.74	4.95
Crosstab nonint	**Model1HSE**	**Model1wt1**	**Model3wt1**	**Model3HSE**
Mean	7.13	6.97	8.89	8.48
Median	6.69	6.46	8.53	7.92
Mode	0.08	0.02	0.01	0.00
Minimum	0.08	0.02	0.01	0.00
Maximum	46.69	41.86	26.63	32.69
Std. Deviation	5.07	4.98	5.59	5.45

Table 35. Complete model comparisons, Cluster 4

Integerised	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	8.26	6.96	6.97	10.52
Median	7.59	4.97	6.23	9.98
Mode	0.00	0.06	0.00	0.02
Minimum	0.00	0.06	0.00	0.02
Maximum	32.28	33.27	34.35	30.27
Std. Deviation	5.92	6.55	5.61	6.68
Nonint HSE wt	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	20.35	7.17	7.60	7.17
Median	20.00	7.13	7.81	7.13
Mode	0.11	0.08	0.12	0.08
Minimum	0.11	0.08	0.12	0.08
Maximum	55.32	17.66	18.47	17.66
Std. Deviation	12.01	4.14	4.46	4.14
Nonint wt 1	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	9.16	6.87	5.93	6.12
Median	8.61	6.69	5.80	5.92
Mode	0.06	0.04	0.02	0.02
Minimum	0.06	0.04	0.02	0.02
Maximum	34.95	21.74	20.16	15.70
Std. Deviation	6.06	4.03	3.78	3.72
Crosstab nonint	**Model1HSE**	**Model1wt1**	**Model3wt1**	**Model3HSE**
Mean	10.82	7.48	13.93	12.87
Median	10.66	6.40	12.88	12.42
Mode	0.05	0.13	0.03	0.06
Minimum	0.05	0.13	0.03	0.06
Maximum	27.66	35.09	43.27	34.16
Std. Deviation	6.68	6.04	9.10	7.75

Table 36. Complete model comparisons, Cluster 5

Integerised	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	15.31	10.75	15.20	6.84
Median	14.38	9.10	14.48	5.51
Mode	13.65	11.30	0.05	0.00
Minimum	0.05	0.00	0.05	0.00
Maximum	48.74	39.64	46.18	41.33
Std. Deviation	9.31	8.20	9.13	5.78
Nonint HSE wt	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	8.46	5.32	5.59	5.32
Median	5.95	4.29	4.40	4.29
Mode	0.01	0.00	0.00	0.00
Minimum	0.01	0.00	0.00	0.00
Maximum	41.27	28.58	30.34	28.58
Std. Deviation	8.54	4.45	4.81	4.45
Nonint wt 1	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Mean	8.99	6.09	6.88	5.59
Median	6.47	5.02	5.68	4.41
Mode	0.01	0.00	0.01	0.00
Minimum	0.01	0.00	0.01	0.00
Maximum	42.80	29.09	35.06	30.41
Std. Deviation	8.93	4.66	5.56	4.78
Crosstab nonint	**Model1HSE**	**Model1wt1**	**Model2wt1**	**Model2HSE**
Mean	7.93	9.22	13.93	12.87
Median	5.78	6.66	12.88	12.42
Mode	0.02	0.02	0.03	0.06
Minimum	0.02	0.02	0.03	0.06
Maximum	36.84	42.51	43.27	34.16
Std. Deviation	7.61	8.98	9.10	7.75

The cross-tabulated data did not improve the model validation as expected. This may be a result of the high specificity of the cross-tabulated constraints; this may have over-constrained the model, and caused it to be too specific so it was unable to estimate unconstrained results. To test this idea, the Model 1 configuration was run for all of the clusters as well as the ‘optimal’ model, however, the results were poorer than for the univariate models (Tables 31-35). The univariate models may lack specificity, however, they appear to be more appropriate to model unconstrained variables, and by extension, diabetes and obesity.

4.6 Optimal model configurations

The optimal model configurations for each cluster are the following:

- Cluster 1: model 2, nonintegerised, initial weight of 1
- Cluster 2: model 3, nonintegerised, initial weight of HSE int_wt
- Cluster 3: model 3, nonintegerised, initial weight of HSE int_wt
- Cluster 4: model 3, nonintegerised, initial weight of 1
- Cluster 5: model 2, nonintegerised, initial weight of HSE int_wt

One concern is that the HSE-defined weights are very high for the 2004 dataset and may not be appropriate for use in the reweighting process. The final decision regarding starting weights (if the HSE weights should be included) depends on whether 2004 data is necessary for inclusion in the final analysis. If the 2004 population is valuable to the dataset by significantly decreasing the percent error for marital status estimations in the optimal models, then the model should be re-run for clusters 2 ,3 and 5 using a starting weight of 1 with the combined dataset to test for a further decrease in the percent error. If the 2004 dataset is not used in the final model the HSE starting weights from 2003 can be maintained for clusters 2, 3 and 5.

The next test run compared the modelled estimations of marital status against the known census distributions for the 2003 population and the combined 2003-4 population. All of the clusters were modelled using the optimal conditions (model configuration and starting weight) specified above. The results varied by cluster, but only Cluster 3 showed a large improvement in the mean percent error using the combined 2003-4 input population (Table 36). Clusters 1, 4 and 5 all had higher percent error using the combined dataset. Cluster 2 had similar levels of error in both models, but higher standard deviation and maximum error using the combined dataset. The higher starting weights for the 2004 dataset did not appear to have a strong effect on the population estimates, as models 2 and 3 had similar or better levels of error compared to the 2003 dataset, and cluster 5 had a higher percent error. Because clusters 1 and 4 (which used a starting weight of 1) also had higher levels of error with the combined dataset, the higher error is not due to the high HSE starting weights.

Table 37. Final model comparisons, unmarried

Cluster 1: Model 2 weight 1	**combined**	**2003 only**
Mean	8.33	7.18
Median	7.01	6.15
Mode	0.02	0.02
Std. Deviation	7.21	5.77
Minimum	0.02	0.02
Maximum	42.72	41.20
Cluster 2: Model 3 HSE weight	**combined**	**2003 only**
Mean	6.23	6.73
Median	4.61	6.30
Mode	0.00	0.00
Std. Deviation	5.44	4.17
Minimum	0.00	0.00
Maximum	31.93	28.77
Cluster 3: Model 3 HSE weight	**combined**	**2003 only**
Mean	5.44	7.20
Median	4.14	7.03
Mode	0.01	0.03
Std. Deviation	4.79	4.54
Minimum	0.01	0.03
Maximum	35.25	34.57
Cluster 4: Model 3 weight 1	**combined**	**2003 only**
Mean	7.97	5.93
Median	7.16	5.80
Mode	0.03	0.02
Std. Deviation	4.75	3.78
Minimum	0.03	0.02
Maximum	22.38	20.16
Cluster 5: Model 2 HSE weight	**combined**	**2003 only**
Mean	10.32	5.32
Median	10.07	4.29
Mode	0.00	0.00
Std. Deviation	5.76	4.45
Minimum	0.00	0.00
Maximum	32.26	28.58

The aim of this research is to provide the best possible approximation of the real-world population. To this end, the best choice is to use the best fitting model to estimate the unknown population characteristics (diabetes, obesity) rather than drawing from a larger micropopulation with higher error for the unconstrained variables. The final version of SimHealth will only include the 2003 dataset because it had the lowest error.

5 Discussion and Conclusion

This paper has set out a series of modifications made to a baseline spatial microsimulation model already used extensively in the UK. Here, the exploration of several modifications allowed for systematic testing and improvement for synthetic population estimation.

The optimal models are unique in several respects. This is the first time that a spatial microsimulation model has been created which can be adjusted for area-specific characteristics, as discussed in section 4.1. The inclusion of a clustering technique to identify areas which are best suited to a specific model configuration is another new approach to creating a more accurate micropopulation (section 4.2). Similarly, the changing of weights from the survey-produced interview weights to the options of universal weights of 1 led to an improvement in some of the areas. Each of these techniques has not previously been introduced into a spatial microsimulation model, and will be the subject of future research applications. The final aspect of SimHealth that lends novelty to this research is the strict specifications for the validation process, as discussed in section 3. The validation using marital status, with low levels of acceptable error, strengthens our confidence in the reliability of the prevalence estimates for obesity and diabetes.

Although integerisation was discarded for all of the final models, this is an area where further investigation is required. The development and adoption of advanced integerisation algorithms may be suitable for later model versions. The cross-tabulation of constraint variables did not improve population estimates, however, this may be a result of the base population size and the small populations within output areas (section 4.5).

This paper is a work in progress, however, the approaches outlined here can lead to improved population estimates for a variety of applications. The next important step is to evaluate how well SimHealth can estimate disease prevalence throughout the study area. Although there is no widespread data on small-area diabetes and obesity prevalence, there is some data available to evaluate model outputs. The specialisation of SimHealth to estimate specific health outcomes by applying the optimal models for each cluster will form the focus of subsequent research.

References

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. doi: 10.1002/psp.351

Ballas, D., Clarke, G., Dorling, D., Rigby, J. and Wheeler, B. (2006). Using Geographical Information Systems and Spatial Microsimulation for the Analysis of Health Inequalities. Health Informatics Journal, 12(1), 65-79.

Ballas, D. and Clarke, G. P. (2001). Modelling the Local Impacts of National Social Policies: A Spatial Microsimulation Approach. Environment and Planning C, 19, 587-606. doi: 10.1068/c0003

Clarke, G. P. and Madden, M. (Eds.) (2001) Regional Science in Business, Springer, Berlin.

Congdon, P. (2006). Estimating Diabetes Prevalence by Small Area in England. Journal of Public Health, 28(1),71-81. doi:10.1093/pubmed/fdi068

Eastern Region Public Health Observatory (ERPHO). (2002). Obesity in the East of England, Eastern Region Public Health Observatory http://www.erpho.org.uk/Download/Public/10334/1/Obesity%20final%20incl%20cover.pdf, accessed on 05/04/07.

Forouhi, N. G., Merrick, D., Goyder, E., Ferguson, B. A., Abbas, J., Lachowycz, K. and Wild, S. H. (2006). Diabetes Prevalence in England, 2001—Estimates from an Epidemiological Model. Diabetic Medicine, 23, 189-197. doi:10.1111/j.1464-5491.2005.01787.x

The Information Centre (2006). Health Survey for England 2004: The Health of Minority Ethnic Groups. http://www.ic.nhs.uk/webfiles/publications/healthsurvey2004ethnicfull/HealthSurveyforEnglandVol1_210406_PDF.pdf, accessed on 10/02/07.

Moon, G., Quarendon, G., Barnard, S., Twigg, L. and Blyth, B. (2007). Fat Nation: Deciphering the Distinctive Geographies of Obesity in England. Social Science & Medicine, 65(1), 20-31. doi:10.1016/j.socscimed.2007.02.046

National Centre for Social Research and University College London. Department of Epidemiology and Public Health, Health Survey for England, 2003 [computer file]. Colchester, Essex: UK Data Archive [distributor], March 2005. SN: 5098.

National Centre for Social Research and University College London. Department of Epidemiology and Public Health, Health Survey for England, 2004 [computer file]. Colchester, Essex: UK Data Archive [distributor], July 2006. SN: 5439.

Norman, P.(1999). Putting iterative proportional fitting on the researcher's desk. Working Paper 03/1999, School of Geography, University of Leeds. http://www.geog.leeds.ac.uk/wpapers/99-3.pdf, accessed on 07/06/07.

Pearce, J., Boyle, P. and Flowerdew, R. (2003). Predicting Smoking Behaviour in Census Output Areas across Scotland. Health and Place, 9, 139-149.

Vickers, D., Rees, P., Birkin, M. (2005). Creating The National Classification Of Census Output Areas: Data, Methods And Results. Working Paper 02/2005, School of Geography, University of Leeds. http://www.geog.leeds.ac.uk/wpapers/05-2.pdf, accessed on 06/06/07.

Voas, D. and Williamson, P. (2000). An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata. International Journal of Population Geography, **6**(5), 349-366.