Submitted to Environment and Planning A 2011-05-26

MoSeS Individual and Household Level Population Data for 2001 Census Output Areas in the UK

Turner A.G.D.

A.G.D.Turner@leeds.ac.uk

Acknowledgements

The UK ESRC funded this work under contract code RES-149-25-0034. MPJ Express was used to parallelise the Java programs to take advantage of multi-processor computer architectures. This allowed results to be generated within a more reasonable time frame. The University of Leeds and the UK National Grid Service supported this work by: providing computational resources on which results were generated; and, expertise to support the development and execution of programs and the management of data on the computational resources. Thanks to NCeSS collaborators, especially those working on MoSeS and to the University of Leeds for your help and support over the years. Special thanks to Graham Clarke for valuable comments on a draft of this paper.

Abstract

Generation of individual and household level population data for 2001 census output areas in the United Kingdom (UK) was work undertaken by the modelling and simulation for e-Social Science (MoSeS) node of the UK National Centre for e-Social Science. There are at least three ways to approach this. One is to synthetically create individual records and households in a way such that their aggregate characteristics match those in census area based aggregate statistics (CAS). Another approach is the focus of this paper, and a third approach could be a mixture of the other ways. This paper details an implementation of a data integration approach which essentially involves selecting sets of records from samples of individual anonymised records to represent the populations of CAS output areas. Comparable area based aggregate statistics generated from the sets of anonymised records are compared with CAS data values. The paper details Genetic Algorithm (GA) optimisation that seek a well fitting set of anonymised records based on goodness of fit measures. Results are presented for 222626 output areas in the UK. Two results are presented with graph based visualisations revealing the goodness of fit of optimisation constraints. These results have both been submitted to the UK data archive and are being made available via the Economic and Social Data Service (ESDS) for others to use as Study Number 6763 (ESDS, 2011). One result uses the Individual Licensed SAR (ISAR) records to represent both the household population and the communal establishment population. The other result uses the Special Licensed Household SAR (HSAR) to represent the household population.

Control constraints were imposed to ensure the results matched some criteria. Optimisation constraints were selected to produce results for applications focussed mainly on the health and social care needs of the dependent population (children, elderly, infirm and otherwise economically inactive).

The programs that generated the results are open source and written in the Java programming language.

1. Introduction

Modelling and Simulation for e-Social Science (MoSeS) is a node of the UK National Centre for e-Social Science funded by the UK Economic and Social Research council from July 2005 until October 2008, (Turner, 2008). The node had a broad remit to develop a community of social science researchers using advanced information and communication technology to develop simulation models for city regions based on individual level demographic models. The models were to be applied for planning and forecasting to explore scenarios for the support of dependent populations. MoSeS was founded on e-Science principles of openness and collaboration and based on the notion that availability of computational resources would not be a major limiting factor, although in practise they always are. The results presented here have been submitted to the UK data archive and are made available via the Economic and Social Data Service (ESDS) for others to use as Study Number 6763 (ESDS, 2011).

The MoSeS Proposal outlined an approach for creating an individual level demographic model, which involved selecting sets of Samples of Census 2001 Anonymised Records (SARs) to represent aggregate Census Area Statistics (CAS) populations for 2001 (Birkin *et al.*, 2004). This initialised population was then to be dynamically modelled or simulated over time. The initial work of integrating the SARs and CAS is the focus of this paper. It builds on the work of Williamson *et al.* (1998) who applied the same concept to integrate similar 1991 UK population census data.

This paper details a Genetic Algorithm (GA) approach to the combinatorial optimisation problem and presents results to encourage their use and analysis. These data can be integrated with other survey data, and used as input data to dynamic UK demographic models.

Some experiments were done to compare GA based results for 2001 with those of an Iterative Proportional Sampling (IPS) program for a limited subset of optimisation variables. The comparison details are not reported here, only the general impression - that results generated by the GA were at least as good, if not better than those produced by the IPS program. Prior to this IPS comparison, results for the GA implementation had already been produced for the UK using a larger number of optimisation constraint variables beyond the handling capabilities of the IPS program at the time.

Section 2 provides a summary of the work that has been done to date. Section 3 provides a general description of the GA and considers the permutations in the optimisation task. Section 4 provides an introduction to the census data used.

Section 5 details the production of Data Set 1 based on an integration of the Individual Licensed SAR (ISAR) and CAS data at Output Area (OA) level for the UK. Section 6 details the production of Data Set 2 which integrates the Special Licensed Household SAR (HSAR) data for representing the household population. The results in Sections 5 and 6 are presented graphically. Section 7 concludes and scopes out further work.

2. MoSeS population initialisation work

MoSeS started in July 2005 and this part of the work (2001 UK Demographic Initialisation) went through two major and many minor iterations during the three year funding period. At each iteration lessons were learned and information was gleaned to refine the process and understand computational requirements. New ways of presenting results were explored and attempts were made to automate the entire process.

The first output for Leeds was produced in December 2005. The program was parallelised and a output for the entire UK was produced by January 2006. These results used a Household Formation Routine (HFR) whereby each household for an output area was formed around individual Household Reference Persons (HRPs), see Birkin *et al. (*2006a) for details. The HFR attempted to assign other individuals to households by matching records using the ISAR variables for age (AGE0), gender (SEX), marital status (MARSTAT), the relationship to the HRP (RELTOHR), family type (FAMTYP), number of dependent children (FNDEPCH), number of elders (HNELDERS), and number of residents (HNRESDNT). The HFR was based on assumptions about the nature of households and was difficult to validate, but it did the important job of grouping individuals into households which was needed for the MoSeS dynamic simulation model being developed.

Progress was presented at the Second International Conference on e-Social Science in July 2006 (Birkin *et al.*, 2006b). Therein the programs that produce individual and household level population data by integrating Census Outputs are referred to as Population Reconstruction Models (PRMs).

Results were presented graphically to MoSeS consultants who judged them to be of insufficient quality. The challenge was to either significantly reduce the difference between variables aggregated from the modelled individual level populations and those in the aggregate census data; or, demonstrate that the GA implemention was reasonable by adapting it to produce results for 1991 census data and compare these with Williamson (2003).

Various things were done in attempts to improve the results. The first modification was to optimise for Household Population (HP) and Communal Establishment Population (CEP) separately. This improved results significantly, but again the results were judged to be of insufficient quality by MoSeS consultants.

Next, results were produced at and analysed for Middle-level Super Output Areas (MSOAs). This was an attempt to both speed up the results generation process and improve the goodness of fit of the optimised results. Despite considerable effort, comparable results could not be generated as rapidly for each MSOA as could be generated for all output areas in each MSOA individually. Also, optimisation at the MSOA level involved a loss of spatial detail in the results which was undesirable. Producing results at MSOA level did help by highlighting a logical flaw in Sampling With No Replacement (SWNR) in producing the results. SWNR constraint meant that for each region, multiple duplicate SAR records were not allowed. Subsequently the algorithms were re-implemented allowing sampling with replacement. These generated significantly better results that were produced quicker using less computational resource.

The next effort to improve results was to experiment with removing the optimisation constraints which measured how well the Household Formation Routine (HFR) assigned individuals to households. The HFR had remained largely untested and it was based on simplifying assumptions about household composition that were probably too general. New results for Leeds and the UK at OA level were generated and graphs were produced for analysis.

At this stage the 2001 Census Special Licensed Household SAR (HSAR) data became available as census outputs. The HSAR data were of considerable interest to the MoSeS consultants and they focussed effort on the development of a new procedure to use these HSAR data maintaining the household groupings to generate the household populations.

HSAR data for England and Wales was available via a special license, but HSAR data for Scotland or Northern Ireland were not readily available. It seemed that there were two choices:

1. Assume that, in general, households in Scotland and Northern Ireland can be represented adequately by those in the HSAR data for England and Wales.
2. Reduce the scope of what was being done and only produce data output for England and Wales rather than the whole UK.

Another issue was that the age variable in the HSAR (AGEH) is coded in 2 year bands up to age 80 (all higher ages are grouped together with a value of 80), whereas the age variable in the ISAR (AGE0) is coded in single years of age up to 16 and from 75 to 95, with 5 groups in between (all higher ages are grouped together with a value of 95). Having tackled this issue, results for the UK at OA level were generated and are compared in this paper.

3. Genetic Algorithm

A Genetic Algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems and has been used in geographical optimisation work since at least 1996 (Diplock and Openshaw, 1996).

A six step GA optimisation algorithm is shown below:

- Step 1. Generate initial results (applying control constraints)
- Step 2. Breed the results to produce more results (applying control constraints)
- Step 3. Measure the goodness of fit of the results (using optimisation constraints)
- Step 4. Select results based on the goodness of fit measure (survival of the fittest)
- Step 5. Repeat Steps 2 to 4 until convergence (or maximum number of iterations is reached)
- Step 6. Store the best fitting result(s)

A search heuristic is needed because the number of possible sets of records from the samples which may form the populations of census output areas is too large for them all to be evaluated. Also, there are 223060 output areas in the UK which is a considerable number of optimisation tasks: If optimisation for one OA takes on average five minutes, then for all OAs, this amounts to a little over two years of processing if the optimisations are done sequentially (one after the other). If all the results are generated simultaneously on a large parallel resource capable of running 223060 processes, the result may be returned in a time a little over the time it takes for the result to be returned for the OA that takes the longest to compute.

Control Constraints (CCs) are variables that have to match for a result to be valid. A CC could be something seemingly simple, e.g. total population. But, just because it is simple to describe, does not mean it is simple to apply and generate valid results. For instance, consider both a total household population CC and a total number of households CC. For the optimisation that selects household records from the HSAR to represent the household population, the issue is one of selecting household records containing an appropriate number of individuals. This is not as straightforward as it might seem. In theory, any variable might be converted into a CC, and a complex set of control constraints may be specified. With more CCs, the results should be more representative in general, but the more complex and detailed the CCs are, the fewer the number of valid results there are, and the harder they are to find.

Optimisation Constraints (OCs) are used to measure the *goodness of fit* of a result. The goodness of fit is a measure of how close the optimised results compare with the optimisation constraints. So, an OC is a variable used in a goodness of fit measure and to evaluate a result in an optimisation process seeking to find results with better goodness of fit. In combining OC measures into a single goodness of fit measure, greater weight may be applied to different OCs. The combination can use various measures of difference. For the results presented here, the weighting for each term was equal and fixed and the combination involved a Normalised Sum of Squared Errors (NSSE) calculated by summing up the squared difference between the CAS count and the SAR aggregate counts, and dividing by one more than the CAS count. (One is added to prevent avoid dividing by zero.)

For the results presented here, the only GA breeding method used is mutation. In breeding, a selection of individuals or households in a population are swapped with others in such a way that CCs are maintained. Allowing the swapping of a large number of individuals or households increases the breadth of search for the genetic algorithm and reduces the likelihood of the optimisation getting stuck with a sub-optimal solution. Yet the amount of swapping in some breeding is small allowing results to fine tune. The GA implementation achieves this by randomising the amount of swapping in breeding to get a range of different amounts of change with each iteration of breeding depending on how the result is converging.

There is an art to GA optimisation and for any particular task, it tends to involve a considerable amount of experimentation to uncover breeding and selection strategies that work well. A couple of things that seemed to work well were: separating results based on their parental lineage so as to keep a variety of solutions (based on original diversity) as opposed to only solutions bred from a single result; also varying parameters in systematic ways rather than pseudo-randomly (or fixing them as unchanging values).

Convergence criteria are used to return good solutions quickly by stopping the algorithm seeking improvement it is unlikely to achieve. If an optimal result has not been found in a reasonable time, or if a specified number of breeding and selection iteration cycles have completed, then halting criteria force the algorithm to return the best results that have been found. In assessing convergence, the number of optimisation iterations that have been completed since a better fitting result is used. If no improvement is found after a given number of iterations, the optimisation may be stopped and a result returned. Convergence parameters can be set and these are used to vary other GA parameters. For instance, difference amounts of mutation are tried, and different numbers of solutions are kept from each optimisation iteration depending on how many optimisation iterations there have been since a better solution was found.

The imposition of different control constraints can have a big effect on how to do breeding. Too many complex control constraints can dramatically slow down the process, whereas adding extra optimisation constraints may have little effect on the speed of the process. With control constraints established, part of the process of setting appropriate GA parameters for breeding involves experimenting with different GA parameters for: the size of initial set of results; selection of results from each iteration; and for convergence criteria.

4. Data sources and output

This section provides some details of 2001 UK Census data focussing on the data outputs integrated in this work. 2001 Census data outputs are produced from the base census data which provides a snapshot of the UK human population collected on paper forms for the enumeration date 2001-04-29. Base census data were digitised and exist as an individual and household level data set for the UK, but in this form they are not readily available for research. If researchers could readily access (even anonymised) yet 'complete' individual and household level census data, then there would be no need for this work which tries to integrate some aggregated and variously generalised census data outputs to provide an estimate of the base census data.

The 2001 Census data outputs integrated in this work are the Census Aggregate/Area Statistics (CAS) and the Individual Licensed and Special Licensed Household Samples of Anonymised Records (ISAR and HSAR respectively). It is important to consider the way these samples and generalisations have been generated from base census data and the limitations this brings. It is not appropriate to provide a full critique of this process here, but suffice to say that the census data outputs are but a shadow of the census data they are derived from and that the values of variables in different outputs are often grouped differently and the results often do not add up as one might logically expect due to disclosure control measures which intentionally introduced errors into the data.

The ISAR is about a 3 per cent sample of individual records for the UK (1843530) and includes variables for age, gender, ethnicity, health, employment status, housing, amenities, family type, geography, social class, education, distance to work, workplace, hours worked and migration. The data are available for England, Wales, Scotland and Northern Ireland. The data are made available for academic research via the Cathie Marsh Centre for Census and Survey Research (CCSR) and more details about them are available on the CCSR and Office for National Statistics (ONS) Web Pages.

The HSAR is about a 1 per cent sample of households for England and Wales (225436) and includes individual records for all individuals assigned to each household (525715 in total). The records are ordered by household and there are identifiers that allow for record linkages between individuals belonging to the same household. There are further variables that (within households) allow for the linkages between family members detailing family relationships. Similar to the ISAR, the HSAR contains variables for age, gender, ethnicity, marital status, social class, education and employment status. It also includes: some household level variables, e.g. housing tenure and number of cars; and, some derived variables, e.g. the number of full time earners in a household, and the age of the youngest dependent child in a household. Details of the HSAR are available on-line on the CCSR and ONS Web Pages. HSAR data are made available for academic research under a special license via the UK Data Archive Economic and Social Data Service (ESDS) as Study Number 5278 (ESDS, 2005).

CAS data are a collection of tables which provide aggregated statistics for Census output areas, the smallest of which is the Output Area (OA). On average an OA contains around 300 people and 100 households. The full set of tables are composed of: CAS tables; CAS Theme Tables - with information on a particular population such as dependent children or pensioner households; and, Univariate tables - which provide a more variable detail for specific variables. Details of these data are publicly available via the ONS Web Pages. CAS data are made available for academic research by the MIMAS Census Dissemination Unit (CDU) and can be accessed via a web interface called CASWEB.

The results submitted to the UK data archive are in American Standard Code for Information Interchange (ASCII) format text files. These essentially list the identifiers of SAR records in two groups for each OA. One group is for Household Population, the other group is for Communal Establishment Population. For those HSAR records only the Household Reference Person (HRP) identifiers are detailed. (Other household members can be looked up using the HRP identifier.) The output data are available from the UK Economic and Social Data Service as Study Number (to be confirmed).

5. Data Set 1: integrating the Individual SAR

This section provides some details of the Genetic Algorithm (GA) implementation for creating a population result for each UK Output Area (OA) that specifies ISAR identifiers for those records used to represent the Household Population (HP) and ISAR identifiers for those records used to represent the Communal Establishment Population (CEP). Control Constraint (CC), Optimisation Constraint (OC) and GA parameter details are provided. Details of the GA parameters used to produce the result are provided. For brevity, only an overview of details and a selection of results are presented here. For more detail see Turner (2011a).

The following CCs were imposed:

1. Counts of HP Household Reference Persons. These are from CAS Table 003 (CAS003) which is divided into HP and CEP counts, and further divided into Male and Female counts for various age groups. The following 4 Age classes are used: 0 to 19; 20 to 29; 30 to 59; and, 60 and over. To constrain to this, use is made of the AGE0, SEX, and RELTOHR variables from the ISAR.
2. Counts of HP. These are from CAS Table 001 (CAS001) which is divided into Male and Female counts for various age groups. The following Age classes are used: 0 to 19; 20 to 29; 30 to 59; and, 60 and over. To constrain to this, use is made of the AGE0 and SEX variables from the ISAR.
3. Counts of CEP. These are from CAS003. The following 25 Age classes are used: 0 to 15 individually, 16 to 19; 20 to 24; 25 to 29; 30 to 44; 45 to 59; 60 to 74; 75 to 84; 85 to 89; and, 90 and over.

The following CAS tables were used to generate optimisation constraints:

1. CAS Key Statistics Table 008 (CASKS008)
2. CAS Key Statistics Table 020 (CASKS020)
3. CAS Key Statistics Table 09b (CASKS09b)
4. CAS Key Statistics Table 09c (CASKS09c)
5. CAS Table 001 (CAS001)
6. CAS Table 002 (CAS002)

CASKS008 are counts of health related variables. The ISAR variables used to compare were: HEALTH; and, LLTI. CASKS020 are counts for household compositions. Counts were grouped for comparison into: Married or Cohabiting Couples with Children; and, Lone Parent Households with Children. The ISAR variable used to compare with these was: FAMTYP. CASKS09b and CASKS09c are counts of the economic activity of Males and Females respectively. Counts for Males and Females were kept distinct and grouped as follows: Unemployed and Age 16 to 24; Economically Active Employed Full Time, Economically Active Employed Part Time, Economically Active Self Employed, Economically Active Unemployed, Economically Inactive Retired, Economically Inactive Permanently Sick or Disabled, Economically Inactive Looking After Home or Family, Economically Inactive Other. The ISAR variables used to compare with these were: AGE0; SEX; and, ECONACT. CAS001 are counts for Males and Females in HP and CEP in various age groups. These were aggregated into total population counts for Males and Females in the following age groups: 0 to 4; 5 to 9; 10 to 14; 15 to 19; 20 to 24; 25 to 29; 30 to 34; 35 to 39; 40 to 44; 45 to 49; 50 to 54; 55 to 59; 60 to 64; 65 to 69; 70 to 74; 75 to 79; 80 to 84; 85 to 89; and, 90 and over. The ISAR variables used to compare with these were: AGE0; and, SEX. CAS002 include counts Males and Females with their marital status. Counts for Males and Females were kept distinct and the following age groups were used: 0 to 15; 16 to 19; 20 to 24; 25 to 29; 30 to 44; 45 to 59; 60 to 64; 65 to 74; 75 to 79; 80 to 84; 85 to 89; and 90 and over. The ISAR variables used to compare with these were: AGE0; SEX; and MARSTAT.

The following GA parameters were specified:

1. Initial population size (InitialPopulationSize)
2. Number of optimisation iterations (NumberOfOptimisationIterations)
3. Maximum number of solutions (MaxNumberOfSolutions)
4. Convergence threshold (ConvergenceThreshold)
5. Maximum number of mutations per child (MaxNumberOfMutationsPerChild)
6. Maximum number of mutation per parent (MaxNumberOfMutationsPerParent)
7. Random seed (RandomSeed)

InitialPopulationSize sets the number of solutions that were initially pseudo randomly created prior to optimisation. NumberOfOptimisationIterations sets the maximum number of optimisation iterations that would be done before a result was output. MaxNumberOfSolutions sets the maximum number of best solutions which would be selected in each optimisation iteration to "survive". ConvergenceThreshold sets the maximum number of iterations after which if no better solution had been found, the best fitting result would be returned. MaxNumberOfMutationsPerChild sets the maximum number of records that could be swapped in a breeding mutation. MaxNumberOfMutationsPerParent sets the maximum number of times a solution would be used in breeding in an optimisation iteration. RandomSeed sets the pseudo random number generator used for the stochastic elements of the heuristic. The pseudo-random number seed is set so that results can be recreated.

For the result submitted to ESDS the run was done in two stages. The first stage was an initialisation with the parameters 1 through 7 set respectively to: 2; 10; 2; 2; 2; 2; and, 0. The second part was based on these initialised results with the parameters 1 through 7 set respectively to: 2; 1000; 2; 2; 2; 2; and, 0

The optimisation constraint graphs shown in Figures 1 to 10 are selected ISAR HP ISAR CEP results referred to in Section 6. Each graph is a plot of 222626 points displayed as crosses, with a Y = X and linear Regression Line. As there are 223060 output areas, then 434 results are not accounted for. The linear regression line equation and coefficient of determination (RSquare) parameters are given in the head of the graph. The variable plotted is given in a key at the foot of the graph. All the CC graphs for each variable are available on-line, Turner (2011b).

Figure 1: OC graph of Males Age 10 to 14

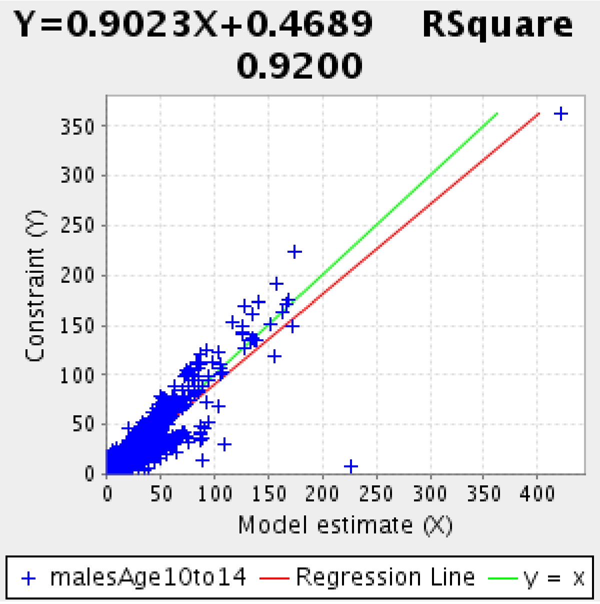


Figure 2: OC graph of Males Age 25 to 29

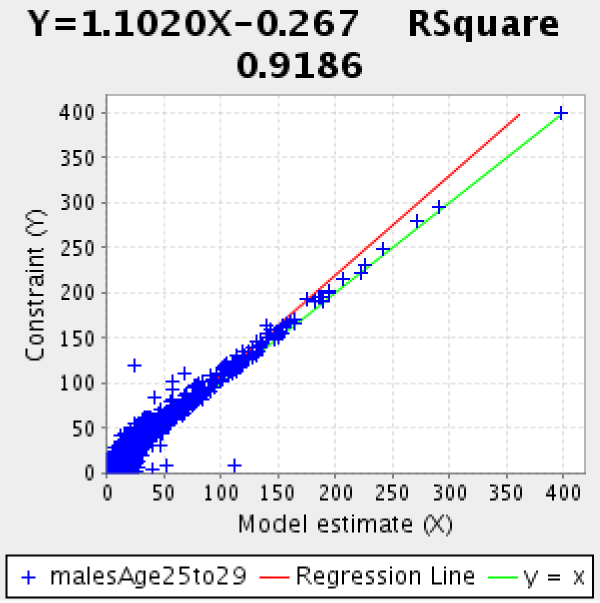


Figure 3: OC graph of Females Age 80 and Over

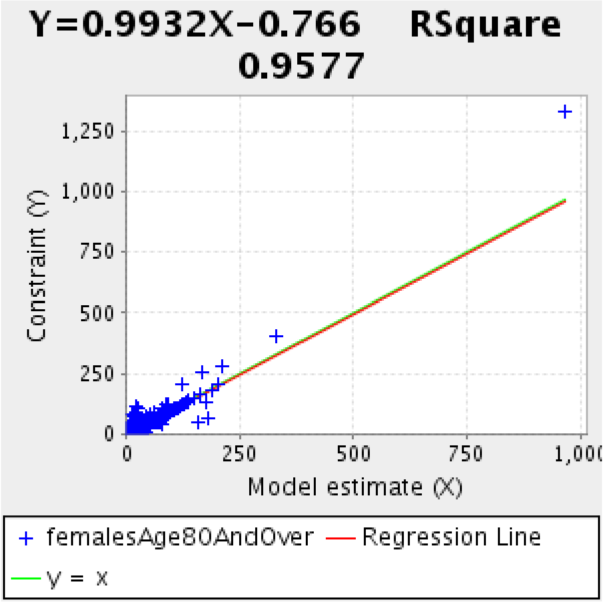


Figure 4: OC graph of Females Married Age 80 and Over

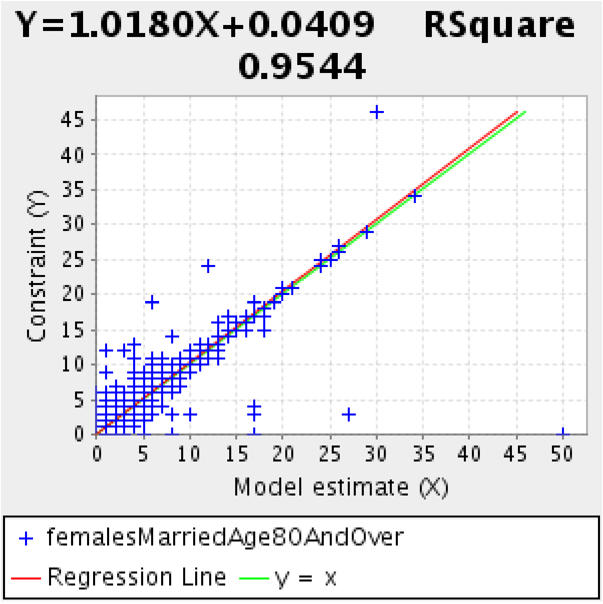


Figure 5: OC graph of People Whose General Health Was Not Good

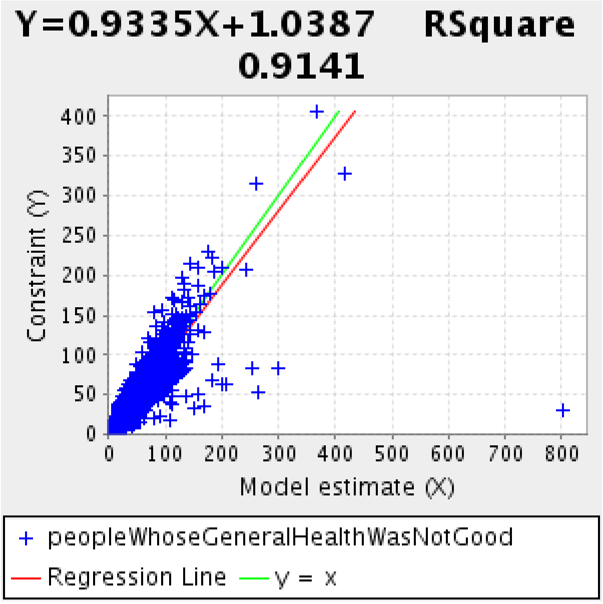


Figure 6: OC graph of People With Limiting Long Term Illness

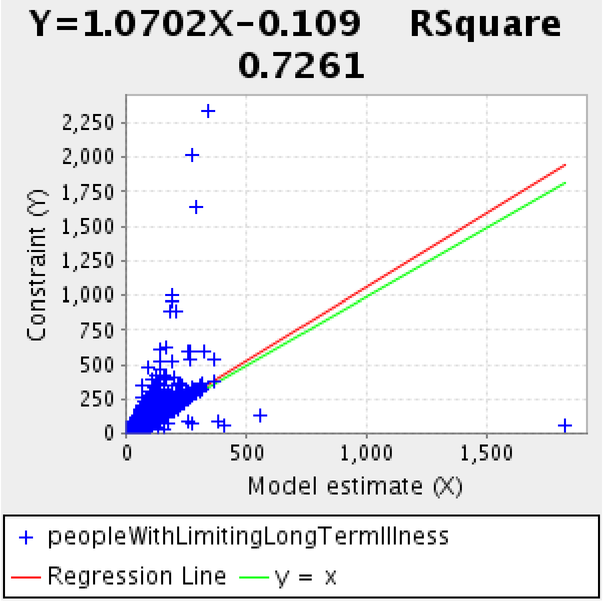


Figure 7: OC graph of Males Age 16 to 74 Economically Inactive Permanently Sick Or Disabled

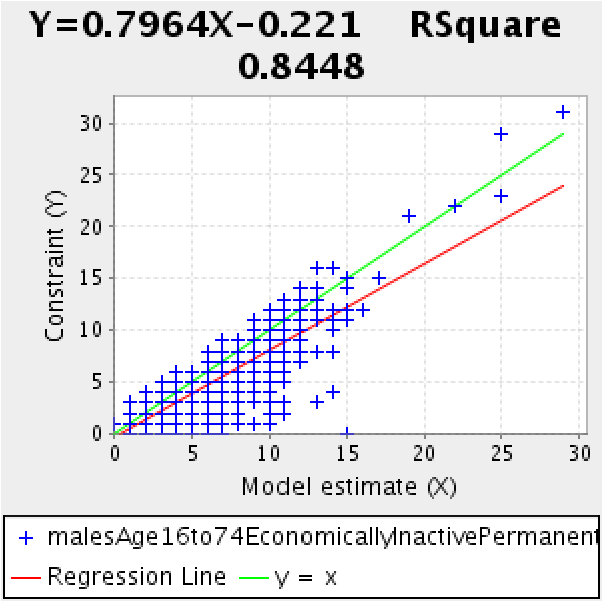


Figure 8: OC graph of Females Age 16 to 74 Economically Active Employed Full Time

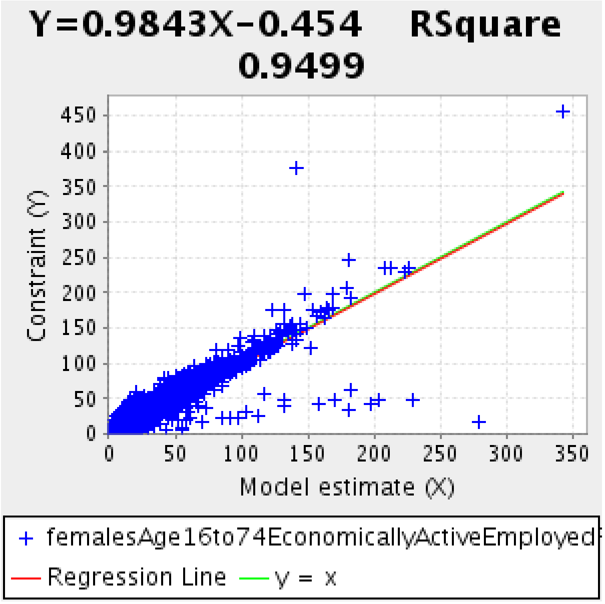


Figure 9: OC graph of Lone Parent Households With Children

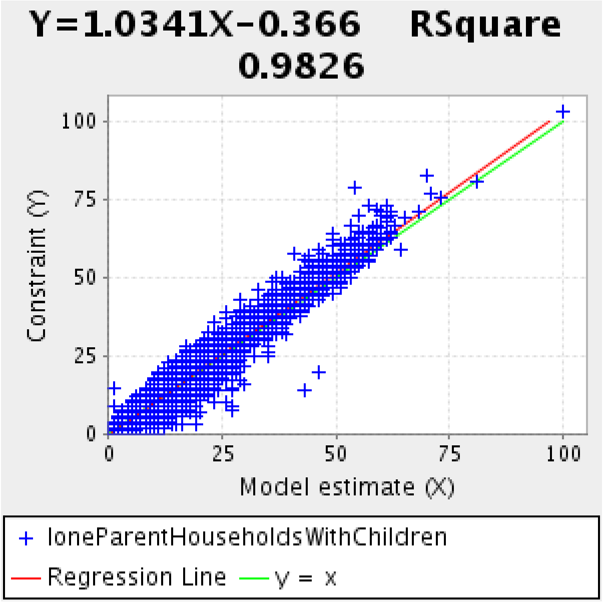
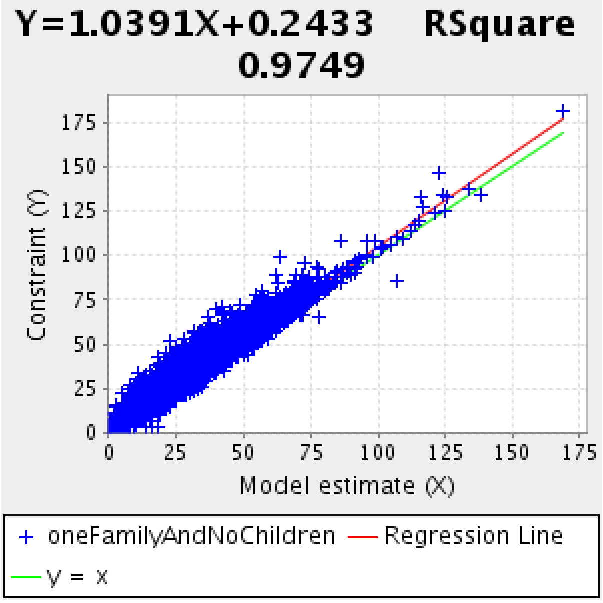


Figure 10: OC graph of One Family and No Children Households



In general the Y = X and linear Regression lines are close and the RSquare values reasonably high. Not all constraint variables have as high RSquare values as those shown in figures 1 to 10. For some variables there are interesting plot characteristics (including groups of outliers) that could be further investigated. It is likely that some output areas have quite unusual populations and these are those that appear as outliers on a number of plots. There could be several reasons for outliers and for poor goodness of fit of some optimisation constraints. Data error cannot be completely ruled out. Further analysis and reprocessing might help improve the results. In particular another approach may be taken for handling special cases.

6. Data Set 2: integrating the Household SAR

This section provides similar details for Data Set 2 as were provided in Section 5 for Data Set 1. For brevity, only an overview of details are provided here, further details are available from Turner (2011).

The same control constraints apply as for Data Set 1 optimisations except that counts of total population by gender (sex) and age are not used. Similar optimisation constraints were also applied as for Data Set 1, but for various reasons there are differences. In the age range 16 to 65, the age variable in the HSAR (AGEH) is more detailed than the ISAR (AGE0) variable, but in other age ranges it is less detailed. The AGEH age ranges are 2 years, yet the age ranges constrained to are mostly in 5 year ranges, so there is a mismatch. Using 10 year age ranges would be less contentious, but risks losing detail.

Additional CAS001 and CAS002 OC variables were created for the age range 30 to 59. Counts of unemployed aged 50 and over for Males and Females based on CASKS09b and CASKS09c were also used as optimisation constraints. It was the in part the greater age variable detail in this age range of HSAR that allowed for this and partly that this was done based on consultation with others. In this it was decided that employment status OC variables were not to be specified in as much detail, so fewer results for these variables have been created.

HSAR aggregates for comparison were derived from HEALTH; LLTI; SEX; ECONACH; AGEH; and, MARSTAH.

Similar GA parameters apply as for ISARHP_ISARCEP optimisations. Similarly, the results submitted to ESDS were created in two stages. The first stage was an initialisation with the parameters 1 through 7 set respectively to: 2; 10; 2; 2; 2; 2; and, 0. The second part was based on these initialised results with the parameters 1 through 7 set respectively to: 2; 1000; 2; 2; 2; 2; and, 0.

The optimisation constraint graphs shown in Figures 11 to 19 are selected HSAR HP ISAR CEP results referred to in Section 7. These have the same characteristics as the graphs shown above. Each graph is a plot of 222626 points. As there are 223060 output areas, then 434 results are not accounted for. All the CC graphs for each variable are available on-line, Turner (2011c).

Figure 11: OC Graph of Males Age 10 to 14

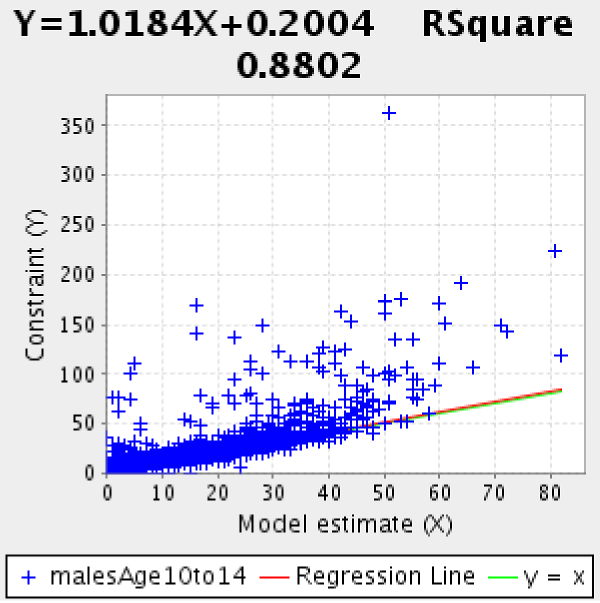


Figure 12: OC Graph of Males Age 25 to 29

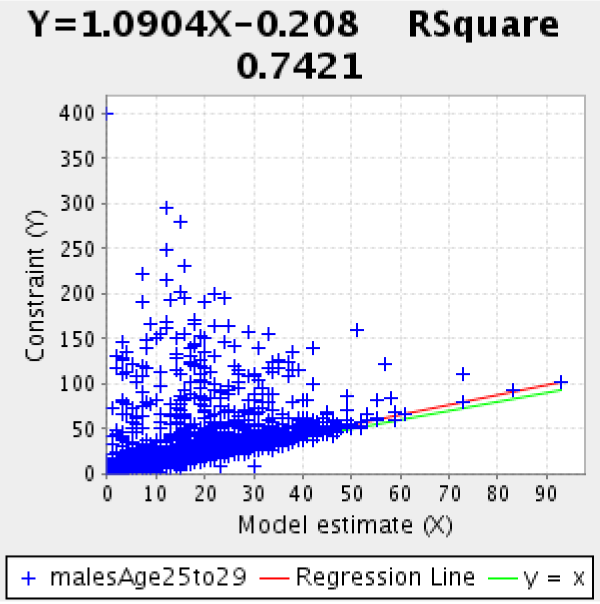


Figure 13: OC Graph of Females Age 80 And Over

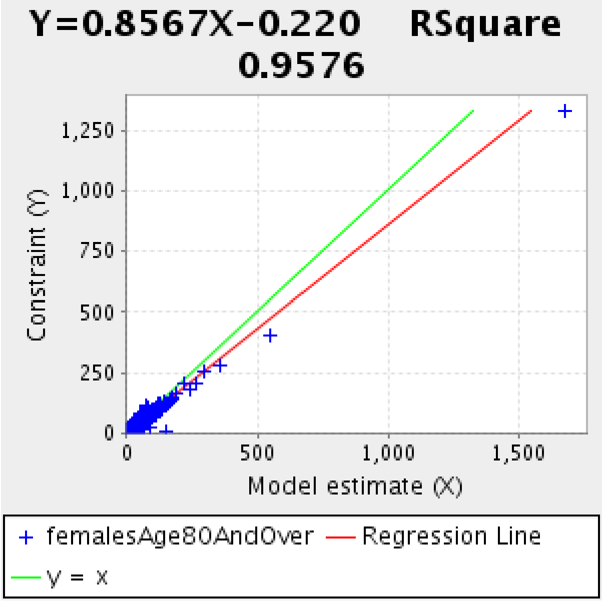


Figure 14: OC Graph of Females Married Age 80 and Over

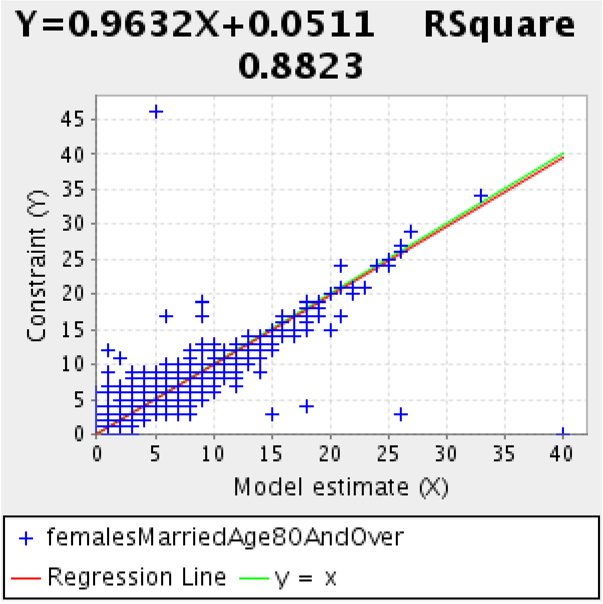


Figure 15: OC Graph of People Whose General Health Was Not Good

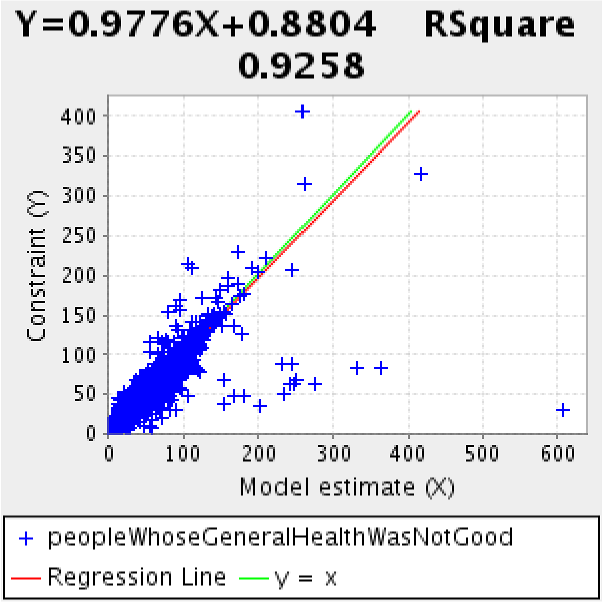


Figure 16: OC Graph of Limiting Long Term Illness

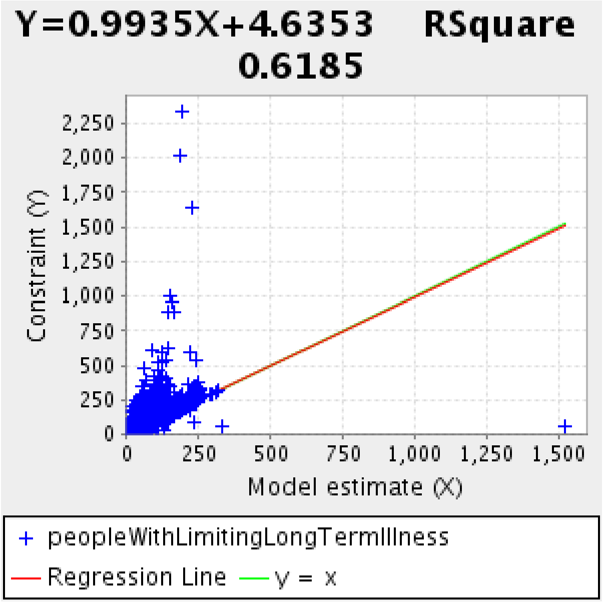


Figure 17: OC Graph of Females Age 16 to 74 Economically Active Employed

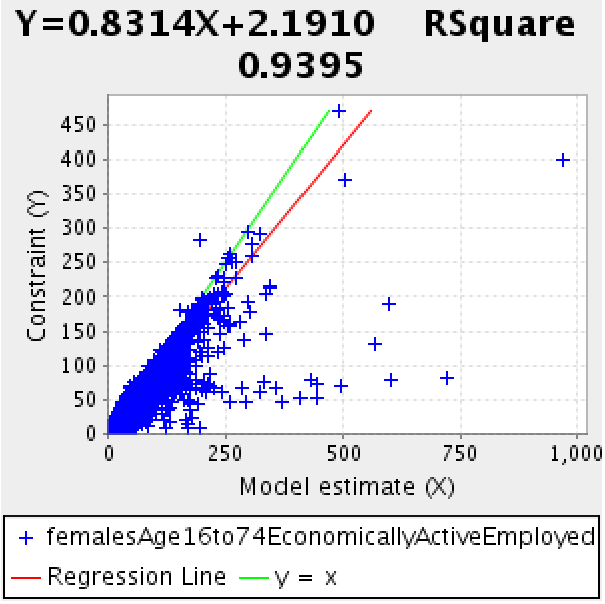


Figure 18: OC Graph of Lone Parent Households With Children

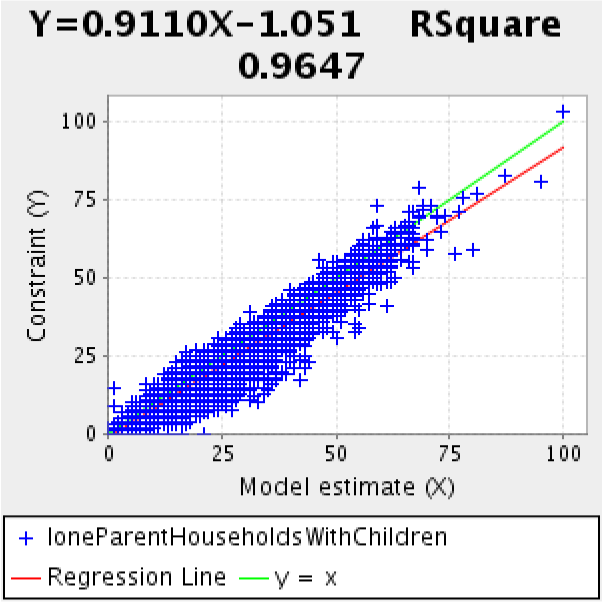
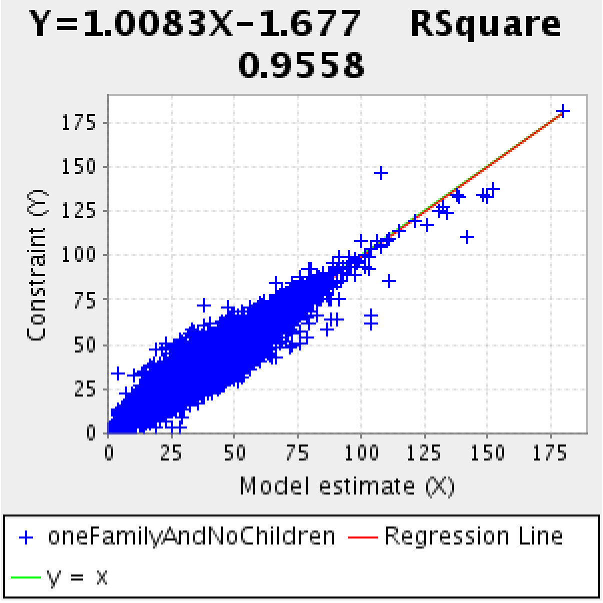


Figure 19: OC Graph of One Family Households with No Children



As with Data Set 1 results, there are reasonably high correlations between the CAS constraints and the aggregated model estimates. In comparison with Data Set 1 results, the model estimates do not in general produce as close a fit with the constraints.

7. Conclusion and scope for further work

Much more can be done to visualise the results and compare these with other data. The results presented here are a few selected graphs which plot census output CAS data on the Y axis against the aggregate SAR results data on the X axis. Such graphs were created for control constraints, optimisation constraints and non-constraint variables. Only optimisation constraint graphs are shown here. Control constraint graphs plotted all points on the Y = X line. Non-constraint graphs revealed that other variables correlated with those used as constraints. All the optimisation constraint graphs and additional graphs are available on-line, see Turner (2011a).

The results presented here can be further analysed by comparing them with other results. The data can be readily accessed by UK academics and used for applications that have the need for an individual level population data for any region of the UK (ESDS, 2011). The programs and which produced the results are also available for others to use and adapt as the see fit (Turner, 2011d). The programs can be compiled and made available as part of a service to be used to produce and share data shaped for specific applications. The 'shaping' may, for example, focus more specifically on young age groups and education; or, middle age groups and economic activity.

The GA approach detailed in this paper could be used to integrate 1991 census data outputs and produce results for comparison with Williamson *et al.* (1998) results which are available on-line, see Williamson (2003).

Finding the optimal solution for a large and diverse combinatorial search with individual 'spikes' of well fitting solutions is difficult. For such a search, any heuristic that effectively assesses pseudo-random combinations has a chance to find the best solution, but unless the search is systematic and covers all possible solutions, there is no way to know that an optimal solution has been found. It is possible to trace what solutions were evaluated, and if this were done, the breadth of search in the solution space could be visualised.

If more complete and more detailed individual and household level census data was more readily available for research, the census data integration work outlined in this paper would be unnecessary. However, attempts to enhance and link census data outputs with data from other sources to develop better demographic models are likely to persist.

References

- Birkin M.H., Dew P., Rees P., Chen H., Clarke M., Keen J., Xu J. (2004) Modelling and Simulation for e-Social Science (MOSES). Proposal submitted to the UK ESRC. http://www.geog.leeds.ac.uk/people/a.turner/projects/MoSeS/documentation/proposal/proposal.doc
- Birkin M.H., Turner A.G.D., Wu B. (2006a) A Synthetic Demographic Model of the UK Population: Methods, Progress and Problems. Paper presented at The Second International Conference on e-Social Science, Manchester, UK. http://www.geog.leeds.ac.uk/people/a.turner/publications/conference/IeSS2006/BirkinSyntheticDemographicModelOfUKPopulation.pdf
- Birkin M.H, Turner A.G.D., Wu B. (2006b) Proof of Concept for a Dynamic Simulation model of the UK Population. Abstract submitted/accepted for The Second International Conference on e-Social Science, Manchester, UK. http://www.geog.leeds.ac.uk/people/a.turner/projects/MoSeS/documentation/articles/BirkinTurnerWu2006b.pdf
- Census.ac.uk Aggregate Statistics Web Page. http://www.census.ac.uk/guides/Area_stats.aspx [Accessed on 2011-04-07]

- Diplock G.J., Openshaw S. (1996) Using simple genetic algorithms to calibrate spatial interaction models. Geographical Analysis, 28, 262-279.
- ESDS (2005) The Economic and Social Data Service Study Number 5278 - 2001 Census: Special Licence Household Sample of Anonymised Records (SL-HSAR) http://www.esds.ac.uk/findingData/snDescription.asp?sn=5278
- ESDS (2011) The Economic and Social Data Service Study Number 6763 - MoSeS http://www.esds.ac.uk/findingData/snDescription.asp?sn=6763
- Huang Z., Williamson P. (2001) A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata. Working Paper October 2001, Department of Geography, University of Liverpool. http://pcwww.liv.ac.uk/~william/microdata/Pop91/Methodology/workingpapers/hw_wp_2001_2.pdf
- ONS Census Area Statistics tables Web Page http://www.statistics.gov.uk/census2001/cas_table_outlines.asp [Accessed on 2011-04-07]
- ONS Samples of Census 2001 Anonymised Records (SARs) Web Page
- http://www.statistics.gov.uk/census2001/cn_117.asp [Accessed on 2011-04-07]
- ONS Home page for the census in England and Wales, includes links to historical information and plans for the 2011 Census. http://www.ons.gov.uk/census/index.html [Accessed on 2011-04-07]
- Turner A.G.D. (2008) Andy Turner's MoSeS Project Web Site. http://www.geog.leeds.ac.uk/people/a.turner/projects/MoSeS/
- Turner (2011a) MoSeS 2001 UK Demographic Initialisation Web Page. http://www.geog.leeds.ac.uk/people/a.turner/projects/MoSeS/documentation/demography/MoSeS2001UKDemographicInitialisation.html

- Turner (2011b) ISAR HP ISAR CEP Optimisation Constraint Graphs http://www.geog.leeds.ac.uk/people/a.turner/projects/MoSeS/documentation/demography/results/2001PopulationInitialisation/NSSE/ISARHP_ISARCEP_2_1000_2_2_2_2_0_1/UK/OptimisationConstraints/OA.xhtml2.0.html
- Turner (2011c) HSAR HP ISAR CEP Optimisation Constraint Graphs http://www.geog.leeds.ac.uk/people/a.turner/projects/MoSeS/documentation/demography/results/2001PopulationInitialisation/NSSE/HSARHP_ISARCEP_2_1000_2_2_2_2_0_1/UK/OptimisationConstraints/OA.xhtml2.0.html
- Turner (2011d) Andy Turner's MoSeS Source Code Web Page http://www.geog.leeds.ac.uk/people/a.turner/src/andyt/java/projects/MoSeS/
- Williamson P. (2003) Small-area synthetic population microdata Web Site http://pcwww.liv.ac.uk/~william/microdata/ [Accessed on 2011-04-07]
- Williamson P., Birkin M.H., Rees, P. (1998) The estimation of population microdata by using data from small area statistics and samples of anonymised records, Environment and Planning A, 30, 785-816.