Modelling and Simulation for e-Social Science: Current Progress

Paul Townend[1], Jie Xu[1], Mark Birkin[2], Andy Turner[2] and Belinda Wu[2]

[1] School of Computing [2] School of Geography
University of Leeds, LS2 9JT, UK
{pt@comp.leeds.ac.uk}

Abstract

This paper reports on the progress that has been made in the Modelling and Simulation for e-Social Science (MoSeS) project at the University of Leeds. This project seeks to use e-Science techniques to develop a national demographic model and simulation of the UK population specified at the level of individuals and households. The architecture of MoSeS is presented and discussed, with particular attention paid to the security requirements faced by the project. The demographic and forecasting models that are used by the project are introduced and discussed, as are the large storage requirements of the project. As part of this discussion, the initial development of a Storage Resource Broker storage facility is presented. The MoSeS portal is then introduced, and initial results of running the demographic and forecasting models are presented and analysed. The paper concludes with a discussion on the future work that will be undertaken as part of this ongoing e-Science project.

1. Introduction

MoSeS (Modelling and Simulation for e-Social Science) is a research node of the *National Centre for e-Social Science* (NCeSS). MoSeS aims to use e-Science techniques to develop a national demographic model and simulation of the UK population specified at the level of individuals and households. The specific aims of the MoSeS project are as follows:

1) To create a flagship modelling and simulation node, in which the capabilities of Grid Computing are mobilised to develop tools whose power and flexibility surpasses existing and previous research outputs.
2) To demonstrate the applicability of grid-enabled modelling and simulation tools within a variety of substantive research and policy environments.
3) To provide a generic framework through which grid-enabled modelling and simulation might be exploited within any problem domain.
4) To encourage the creation of a community of social scientists and policy users with a shared interest in modelling and simulation for e-social science problems.

There are an abundance of simulation games relating to people, cities and societies (past, present and future). We pose the question of what would be the impact of transferring these simulations into a real world environment. Our specific interest is in cities and regions, with an aim of building simulation models of interactions between individuals, groups or neighbourhoods within large metropolitan areas. Such simulations can form the basis of a wide range of applications in both *e-Research* and *public policy analysis*, with potentially substantial benefits such as:

1) A big policy impact through the generation of effective predictions.
2) A potential 'wind tunnel' or 'flight simulator' analogy: planners can gauge the effects of development scenarios in a laboratory environment.
3) The use of simulations as a pedagogic tool allows planners to refine understanding of systemic behaviour and alternative futures, thus aiding clarity of thinking and improved decision-making.

Specifically, MoSeS aims to develop scenarios in the domains of health, transport, business. For example, one health scenario would be to provide perspectives on medical and social care within local communities for a dynamic and ageing population. A scenario in the transport domain might concern the sustainability of

transport networks in response to demographic change and economic restructuring: for example, what kind of transport network is capable of sustaining long-term economic growth in West Yorkshire, Greater Manchester, and the intervening areas – the 'Northern Way'. A scenario in the business domain might include the impact of diurnal population movements on retail location and profitability; or the impacts of a changing retirement age on personal wealth and living standards.

The results of the simulations and scenarios produced by MoSeS can be accessed and queried by a user through a JSR-168 compliant portlet interface running on top of the Gridsphere portlet engine, with visualisation of results provided by both graphing libraries and also a Google Maps interface.

The MoSeS project stands to benefit from e-Science technologies in a number of ways; in particular, the simulation model will draw on diverse, virtualised data sources, will deploy models which are richly specific and therefore computationally intensive, and will provide outputs to a spatially distributed community of researchers and policy-makers. MoSeS is building relationships with policy users in Social Services, Health Care Trusts, urban planning, consultancy and other domains in order to demonstrate the viability and potential impact of simulation modelling, enabled by e-Science.

This paper describes the progress of the project to date in implementing the baseline demographic model. In Section 2 of the paper we discuss and evaluate the major areas of work within the project – namely the creation of demographic and forecasting models, the storage resource broker based storage system, and the JSR-168 compliant portlets that form the user interface to the MoSeS system. We then introduce some initial results that have been produced by the project, and present an initial analysis and evaluation of the effectiveness of the existing models. The paper concludes with a discussion on the future directions that the MoSeS project plans to take.

2. MoSeS Architecture

Figure 1 shows the current architecture of the MoSeS project. There are three major components to MoSeS; computationally intensive demographic and forecasting modelling, virtualised storage resources brokered through the use of a Storage Resource Broker (SRB) [RAJ03] cluster, and the collection of JSR-168 compliant portlets that make up the user interface to the project.

This architecture is very much designed to take into account the distributed and decentralised nature of e-Science research; for example, the data resources used as part of the MoSeS project are distributed both spatially and across organisational boundaries.

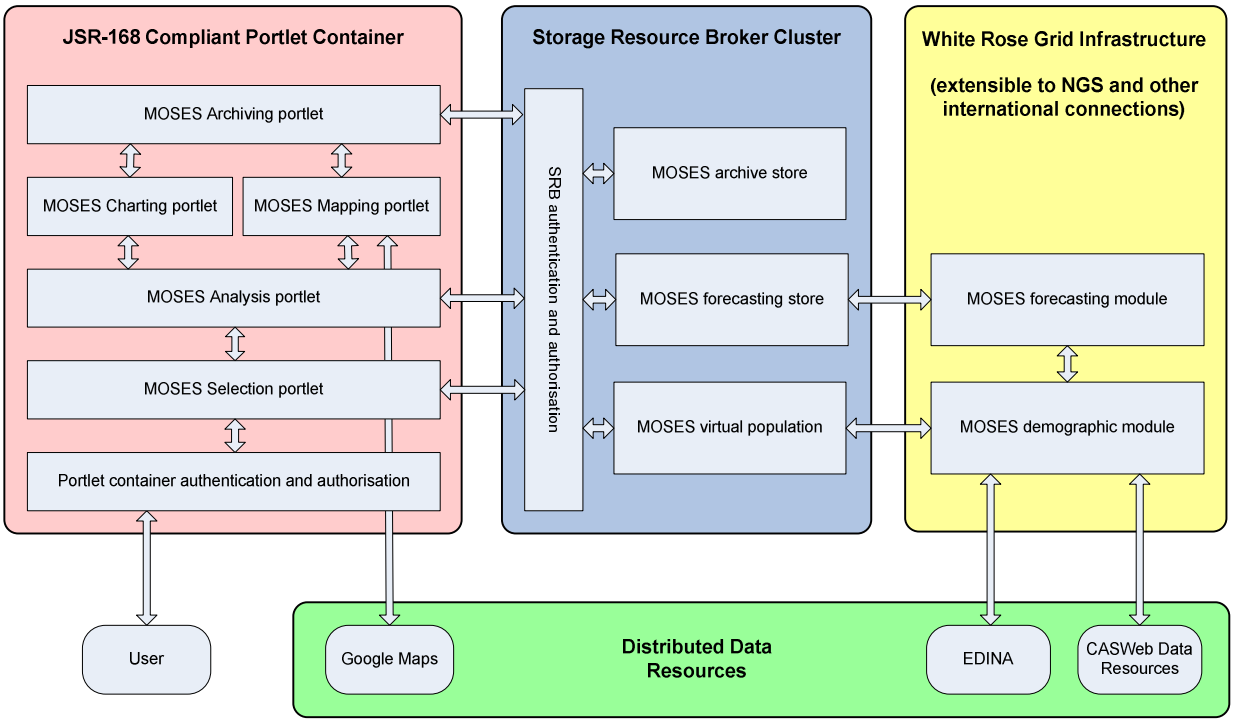

Figure 1. MoSeS Architecture

Likewise, the computational requirements of the MoSeS forecasting and demographic modules mean that processing must be performed in a distributed way; this is currently achieved through the use of a Beowulf cluster at the University of Leeds, but is intended to be scaled up for use on a larger Grid system, such as the UK National Grid Service (NGS). The extremely large storage facilities that MoSeS requires are handled by a distributed Storage Resource Broker resource, and the user interface to MoSeS is achieved through the development of a set of specific portlets which can be grouped according to the needs of each individual user. Visualisation of results is in part achieved through interactions with third party mapping software (currently Google Maps).

Such a decentralised and distributed architecture brings with it a number of challenges, particularly within the social-science domain. Of these, the most important is that of data security; MoSeS uses often highly confidential mapping, health and census data, and it is imperative that such data is not accessible by a user without the necessary authorisation. Current access control mechanisms are as follows:

1) **Portlet Container**. The MoSeS project currently uses the Gridsphere portlet container; users are required to log in to the portal using standard Gridsphere authentication mechanisms. Individual portlets are configured to work on a role-based manner, with different Gridsphere users being manually assigned different roles during the creation of their account.

2) **Storage resource broker cluster**. The SRB cluster used by the MoSeS project stores both map data and also the results of processing performed by the demographic and forecasting modules. Data is protected through the use of SRB authentication and authorisation processes; this is again currently a manual process, with different SRB users assigned access rights to specific datasets.

The security requirements brought about by the processing of such confidential data are also of concern. The computational demands of MoSeS are high, and distributed processing is necessary to achieve results within a reasonable timeframe; however, staging confidential data onto third party, potentially untrusted nodes is a great concern. Currently, only clusters of machines that are internal to the University of Leeds are being used for computational tasks; novel security mechanisms need to be investigated before the computational aspect of MoSeS can be extended outside of the University of Leeds.

There now follows a discussion on each major component of the MoSeS architecture .

2.1 Demographic model

Many social science microsimulation models use the principle of synthetic estimation, whereby a model is used to generate individual characteristics on a progressive basis using compound probabilities. For example, suppose that an individual aged 55 works as a printer and lives in a suburb of Leeds with a wife and no other dependents. What is the probability that such an individual suffers from a limiting long-term illness (LLTI)? Such probabilities can be extracted from UK census data and used as a basis for assigning health status to such a synthetic individual [BIR88]. This process is performed in order to create a *virtual population*. Once this virtual population has been created, forecasting models can be used to move the population forward in time, in order to analyse different health, transport and business scenarios.

The Moses Population Recreation Model (PRM) uses the principle of 'reweighting' to create a virtual population. We already have a distribution of individuals (through analysis of Individual Samples of Anonymised Records – ISARs – provided by the UK census), but these are not representative of each census Output Area (OA). So we need to select from the ISAR in order to define a subset which is more representative. This can be thought of as assigning weights of one or zero to each ISAR record. In order to reweight data from the UK SARs, the most common heuristic has been simulated annealing, e.g. [WIL98, BAL04].

The performance of this method has been good, although it has not been benchmarked explicitly against other approaches. The major drawback of simulated annealing is that it requires significant computation, particularly in the context of a national simulation involving more than two hundred thousand small areas. The method can potentially be made parallel across areas but not within areas. In other words, within a multi-processor environment, we could send data for different areas to a number of different processors, but it would make no sense to send data for a single area to more than one processor.

An alternative method to simulated annealing is a genetic algorithm (GA), as suggested by Williamson et al (1998). Intuition suggests that a GA should be well-suited to an optimisation in which zero-one weights are to be applied to a database (a natural gene string). One would expect the main drawback of such an approach to be its computational intensity, but this process is easily parallelised both within and between

areas. In other words, the solution method will involve the creation and evolution of a candidate population of solutions for each area, and there is no reason why each candidate solution need not be assigned to a different processor within a multi-processor environment.

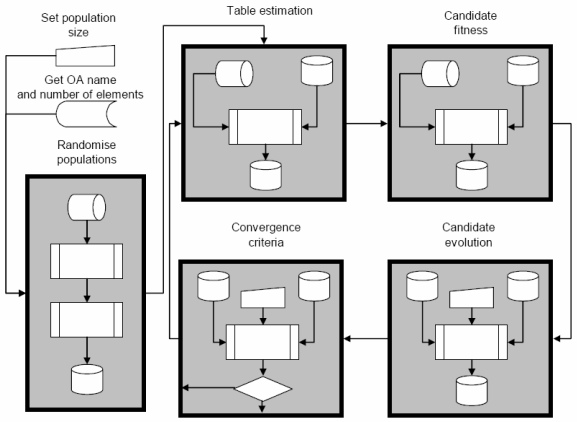

Figure 2. MoSeS PRM Architecture

The results of some experiments with a GA implementation of the demographic model are reported in this paper in section 3.1; the architecture for the MoSeS PRM is shown in figure 2.

2.2 Forecasting model

The first stage towards the development of a forecasting model for the MoSeS has been the instantiation of a 'toy model'. In simple terms, the objective of the toy model is to include such detail only as is needed to demonstrate the utility and general principles behind the modelling approach.

The spatial extent of the toy model is confined to the city of Leeds, which is treated as a closed system. Within the city, individual people and household groups are assigned to precise individual locations. The starting point of the simulation is the virtual population constructed through the demographic modelling process detailed previously. The objectives of the simulation are twofold:

1) to develop a process of household formation through which individuals are combined into household and family structures.
2) Secondly, a process of household dynamics, through which households are aged, fragment and dissolve.

The process of household formation is shown in the form of a flow diagram, illustrated in figure 3. Note that what is represented here is a static or synthetic process: we are trying to estimate the relationships between individuals within households within a baseline population; rather than, for example, modelling dynamic processes by which single individuals choose their partners, marry and have families. Sample results from the forecasting model are given in section 3.2.

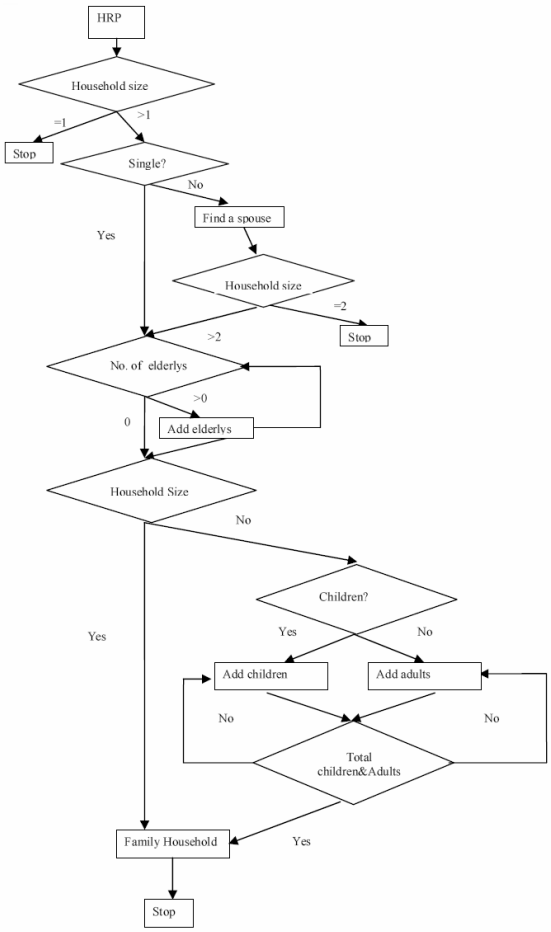

Figure 3. The Household Formation Process

2.3 Storage Resource Broker Cluster

The storage demands created by the MoSeS project are extremely large; given the large amounts of computing power (and time) required to generate results, it is important that results can be stored at a high level of

granularity. However, when considering a UK-wide simulation comprising of 60 million people, a 30 year archive of results (taken with one snapshot per year) may consume as much as 3 terabytes of data.

For this reason, we have begun to deploy and integrate a distributed storage network implementing Storage Resource Broker technology. This allows data to be stored across multiple nodes and accessed in a uniform, location and platform agnostic manner by all MoSeS applications. At the current stage in the project, not all archival data is currently stored in this resource, however, due to the large network demands placed on transferring such large volumes of data; it may be the case in future that the SRB network is only used for data that has been specifically flagged as long-term archival.

2.4 JSR-168 Compliant Portlet Interface

In order to allow users – both expert and novice – to interact with MoSeS in order to perform experimentation, and analyse and visualise results, a series of JSR-168 compliant portlets have been developed. Portlets are pluggable user interface components that are managed and displayed in a web portal, and function by producing fragments of markup code (usually HTML) that are aggregated into a portal page. JSR-168 refers to the Java Portlet Specification, designed to define a contract between a portlet container and portlets, as well as to provide a convenient programming model for portlet developers.

In our case, we have developed a MoSeS portal page using the Gridsphere [NOV04] portal framework, thus allowing any user with a web browser and an account to be able to interact with MoSeS. Each individual MoSeS portlet can be plugged into this portal to form functioning applications, depending on the role of the user; a screenshot of the MoSeS portal, running a number of portlets, is shown in figure 4.

The portlets currently developed are as shown in the architecture diagram in figure 1; the *selection portlet* (as shown in figure 4) allows a user to choose a UK city from a pre-defined list, and select areas from it to analyse. The *analysis portlet* allows a user to invoke an analysis of the simulation results (either stored locally or on the SRB resource) that have been produced for that area; this analysis is mostly concerned with aggregating data from each individual agent record that the simulation has produced. The *charting portlet* is concerned with producing graphs and charts (such as pie charts, waterfall charts, etc.) of these aggregated results in order to produce comparisons such as "Age vs. Health" and "LLTI vs. Ethnicity" for the selected areas.

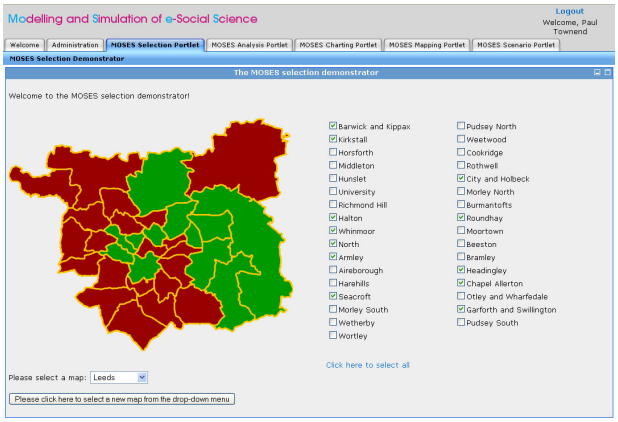

Figure 4. Screenshot of the MoSeS portal

The *mapping portlet* is concerned with producing shaded maps of the selected region, based on user-selected values. For example, a map may show populations per area, average car ownership per area, percentage of population with Diabetes per area, etc. This portlet has recently been extended to provide a Google Maps interface, a screenshot of which is shown in figure 5. This interface has in part been developed using "Google Map Creator" technology provided by [GOO07] as part of the GeoVUE project.

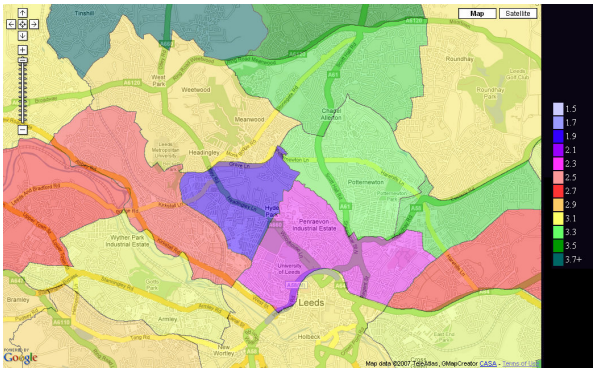

Figure 5. MoSeS Google Maps Interface showing Diabetes incidence per Ward in Leeds

Finally, the *scenario* portlet is concerned with allowing users to perform "what if" scenarios in the selected areas. Currently, the only available scenario is the "Health" domain. Users are able to analyse proportions of the selected population with chest problems, blood problems, heart problems, diabetes, etc. under a number of different assumptions (such as the "baseline forecast", the "healthy" forecast and the "very healthy forecast" which alter assumptions about the lifestyle of agents). These scenarios can be rolled

forward by a user-selected number of years in order to assess long-term trends.

3. Initial results

The work being performed on the MoSeS demographic and forecasting models is still ongoing, but a number of initial results have been produced.

3.1 Demographic model results

In order to test the effectiveness of the MoSeS demographic model, detailed testing has been performed on models of the Leeds Local Authority District (LAD). Our main model performance criterion has been to produce spatial distributions of a series of demographic attributes, and to compare modelled data to known distributions. These comparisons have been performed at a ward level within the Leeds LAD using an Index of Dissimilarity. Results are shown at Table 1; the major themes emerging from these results appear to be as follows:

1) The relationship between controls, optimising and co-varying attributes is as expected, with closest adherence for constraints and the loosest for co-variation
2) Even the control attributes are not distributed perfectly, which could be a facet of some of the data issues described earlier
3) Co-variation does not appear to be very effectively represented within the algorithm at present. For example, ethnic status is not as strongly spatially clustered within the model as in reality. This probably reflects the fact that none of the controls or optimising attributes currently selected is closely related to ethnic status. An interesting question is what controls and optimising attributes might be selected in order to give the best overall profile of a city and its constituent neighbourhoods.

3.2 Forecasting model results

In terms of the forecasting model, sample results from the household formation process are shown in Table 2. One of the objectives of the simulation at this point is to 'find' a spouse for each of the Household Representative Persons (HRP) within the base population. Ideally, therefore, we expect the number of households formed (right hand column) to match the number of HRPs (left-hand column). Although the rate of household formation approaches 100% in some cases, this is never achieved in any of the examples shown in Table 2, and in some cases as many as 20% of households are unformed.

At the time of writing, this problem appears to be attributable to a lack of 'spouse candidates' within the base population. Steps are therefore in process to regenerate the base population to allow a more satisfactory set of households to be created. The dynamics of household change are shown in Figure 6.

At this stage, we are particularly conscious of the important impacts of changing household structures on health care services. Thus we have a particular interest in changing health status over time, and the idea that as one partner within an elderly couple becomes infirm, then one or both household members may move from independence into care. As this model becomes more sophisticated, we hope to represent more complex interactions, such as the importance of local service provision, and even the strength of both social and family networks within the care process.

Table 1. Initial MoSeS Demographic Model results for Leeds LAD.

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

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

Table 2. Forecasting model results for Household formation.

	Population Structure by OA				spouse candidates		
OA	Total	HRP	non-HRP	children	singleNonHRP Adults(include Elderly)	marriedNonHRP Adults(include elderly)	Household Formed
00DAFA0037	221	114	107	19	80	8	96
00DAFA0001	304	143	161	59	95	7	132
00DAFC0023	229	94	135	50	77	8	77
00DAFD0028	296	139	157	55	90	12	122
00DAFH0070	298	154	144	68	73	3	98
00DAFK0075	326	114	212	94	104	14	110
00DAFR0058	310	128	182	46	117	19	119
00DAFL0039	309	122	187	58	113	16	114
00DAFS0065	245	103	142	49	85	8	97
00DAFX0025	319	117	202	64	128	10	115
00DAGA0053	221	101	120	50	56	14	78
00DAGC0012	253	129	124	50	69	5	101

Currently, however, the forecasting model is a much more stylistic and simplified proof of concept. Households and their members are allowed to age, dissolve, and die without being regenerated through new household formation. Therefore as the model is run forward over time, the populations within each area gradually wither and disappear (see Figure 7). In areas with a relatively elderly demographic structure – such as FL39 and FY72 – the rate of decline is much swifter than in younger areas such as FA37 and FC23.

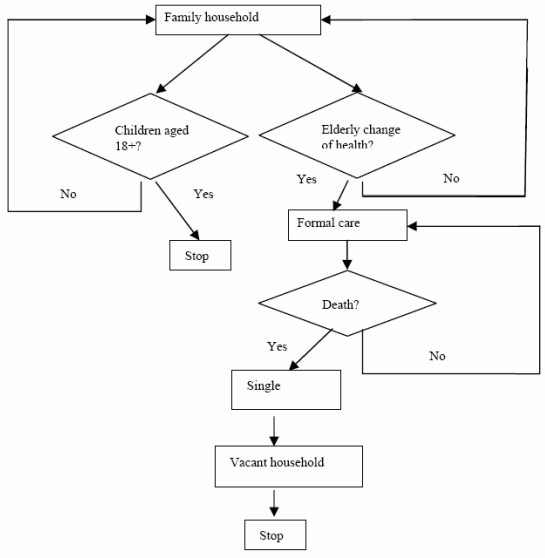

Figure 6. Dynamics of household change

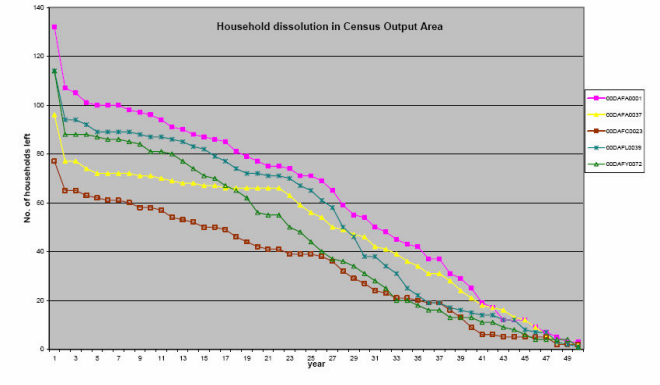

Figure 7. Population over time

4. Conclusions and Future Work

The MoSeS project seeks to use e-Science techniques to develop a national demographic model and simulation of the UK population specified at the level of individuals and households. A large amount of implementational work has now been performed, including the creation of demographic and forecasting models and an extensive portlet-based user interface that combines with Google Maps to create an impressive and easy-to-use user experience. Initial results have been recorded for both models used by the project, and as a result of analysis performed on these results, a number of issues and challenges are actively being addressed.

Future work consists of continuing to improve the portlet interface, in addition to further refining the

simulation models. A particular area of concern are the security requirements, discussed in section 2, that arise from both hosting and processing confidential UK Census, health and map data – not only in terms of the raw data but also the ensuing forecasts (for example, it may be possible to reverse engineer a forecast in order to obtain the original data). Currently, the in-built security authentication and authorisation mechanisms in SRB and Gridsphere are used, but it is necessary to further investigate whether these systems are sufficient for the needs of the MoSeS project. Furthermore, a detailed investigation into the feasibility and security of running MoSeS computational jobs on nodes external to the University of Leeds needs to be performed before the computational aspect of the project can be rolled out further.

5. References

[BAL04] D. Ballas, D. Rossiter, B. Thomas, G.P. Clarke, and D. Dorling, "Geography matters: simulating the local impacts of national social policies", Joseph Rowntree Foundation, York, 2004.

[BEC96] RJ Beckmann, K Baggerly, M McKay, "Creating synthetic baseline populations", in *Transportation Research A, 30*, 415-429, 1996.

[BIR88] M Birkin and M Clarke, "SYNTHESIS: A SYNTHetic Spatial Information System for urban modelling and spatial planning" in *Environment and Planning A, 20*,1645-1671, 1988.

[GOO07] Google Map Creator, UCL Centre for Advanced Spatial Analysis, http://www.casa.ucl.ac.uk/software/googlemapcreator.asp

[NOV04] J. Novotny, M. Russell, O. Wehrens, "GridSphere: a portal framework for building collaborations", in Concurrency and Computation: Practice and Experience, Vol. 16, No. 5, March 2004.

[RAJ03] A. Rajasekar, M. Wan, R. Moore, W. Schroeder, G. Kremenek, A. Jagatheesan, C. Cowart, B. Zhu, S.-Y. Chen, and R. Olschanowsky, "Storage Resource Broker—Managing Distributed Data in a Grid," Computer Society of India Journal, Special Issue on SAN 33, No. 4, 42–54 (October 2003).

[REE05] P Rees, J Parsons. and P Norman, "Making an estimate of the number of people and households for output areas in the 2001 Census", in *Population Trends, Winter 2005*.

[WIL98] P. Williamson, M. Birkin, P. Rees, "The estimation of population microdata by using data from small area statistics and samples of anonymised records", in Environment and Planning A, 30, 785-816, 1998.