

Modelling and Simulation for e-Social Science

Paul Townend¹, Jie Xu¹, Mark Birkin², Andy Turner², Belinda Wu²

¹School of Computing, University of Leeds, LS2 9JT

²School of Geography, University of Leeds, LS2 9JT

pt@comp.leeds.ac.uk

Summary

MoSeS (*Modelling and Simulation for e-Social Science*) is a research node of the National Centre for e-Social Science (NCeSS). MoSeS uses e-Science techniques to execute an events-driven model which simulates discrete demographic processes; this allows us to project the UK population 25 years into the future. This paper describes the architecture, simulation methodology, and latest results obtained by MoSeS.

Key words: *modelling, simulation, service-oriented, e-science*

1. Introduction

MoSeS (*Modelling and Simulation for e-Social Science*) is a research node of the National Centre for e-Social Science (NCeSS). MoSeS uses e-Science techniques to execute an events-driven model which simulates discrete demographic processes; this allows us to project the UK population 25 years into the future. Whilst this approach is grounded in the methods of micro-simulation, concepts from spatial interaction modelling and agent-based systems are incorporated in an innovative way. 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). The underlying aim of the research reported in this paper is to translate such games into real world policy environments. If planners were equipped with the means (through simulation) to understand social and demographic changes in response to shifts in policy, such a device would have valuable practical applications as both a ‘decision support system’, and as a pedagogic tool for understanding how cities work. As an academic and intellectual challenge, the ability to reproduce and predict the behaviour of real city systems constitutes the ultimate

demonstration of a deep understanding of such systems. We wish to produce a simulation model of the UK population, as it now is and as it can be expected to develop over a twenty-five year time horizon. 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 changing 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 MoSeS project stands to benefit from e-Science technologies in a number of ways; in particular, the simulation model draws on diverse data sources, deploys models which are richly specific and therefore computationally intensive, and provides 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. An e-science approach such as this will enable a clear advance over existing techniques, as it will greatly ease the integration of new datasets, and quickly take advantage of new and large-scale computing resources in a dynamic fashion.

In order to exploit the full potential of e-Social Science technologies, flexible and advanced interfaces to MoSeS have been created; in addition to describing MoSeS simulations and latest results, this paper explores the capabilities offered by exposing MoSeS functionality through *service-orientation* and *Web 2.0* technologies. Service-orientation is emerging as a highly useful means of developing flexible, agile, and dependable software systems. A service can be defined as “*a mechanism to enable access to a set of one or more capabilities, where the access is provided using a prescribed interface and is exercised consistent with constraints and policies as specified by the service description.*”, whilst a service-oriented architecture can be defined as “*an application architecture within which all functions are defined as independent services with well-defined invocable interfaces, which can be called in defined sequences to form business processes.*”

Section 2 discusses the service-oriented MoSeS architecture, whilst Section 3 gives an overview of how MoSeS simulations are performed. Section 4 presents the latest results obtained by MoSeS simulation runs on the city of Leeds, UK, and Section 5 describes the Web 2.0 technology that has been integrated into MoSeS to provide richer visualisations of simulations and their results. This paper is concluded in Section 6.

2. MoSeS Architecture

To achieve the aims detailed in section 1, the software architecture employed by the MoSeS project needs to be capable of securely storing large quantities of simulation data, dynamically retrieving diverse data from spatially distributed resources, utilising the capabilities of high performance computing (HPC) resources, and visualising the results of simulation runs. In addition to this, a high level of flexibility is very much desired, to enable MoSeS software to be quickly adapted to utilise new datasets, cope with changes in the structure of existing datasets, perform simulations based on new attributes, etc. In order to address these requirements, it was decided to *service-enable* MoSeS functionality. Service-orientation aims to facilitate the development of complex, dynamic and interorganisational systems. It is also designed to greatly simplify the process of integrating existing legacy systems, and has a profound impact on the software development process. By implementing MoSeS functionality as a set of invocable services, a number of advantages are introduced. These can be summarised as follows:

- 1) *Multiple entry points to the system.* By exposing MoSeS functionality as a set of invocable services, it becomes possible to quickly create new user interfaces to MoSeS. The functionality of each service can be utilised by programs written in any language, and GUI clients can be created as front ends to both workflows (which can be defined as the description of the sequence of services that must be invoked to form a given scientific process) and individual services. Another important aspect of having multiple entry points into the systems is that other e-Social Science applications can now integrate with MoSeS much more easily, simply by invoking the required service.
- 2) *Richer user interfaces.* Multiple user interfaces with more features can be created in order to improve the user experience. For example, a Java application can be written as a front end to MoSeS, utilising the same service functionality that a JSR-168 compliant Portlet interface uses, but with the addition of interactive maps, due to the Java GUI not being constrained by JSR-168 standards, etc.
- 3) *Increased user flexibility.* In addition to pre-written analyses, users can now construct new MoSeS workflows through the use of high-level workflow engines such as Taverna and ActiveBPEL. This enables users to develop, share and re-use routines for producing specific maps, charts, and reports. Developing such workflows is a far more intuitive form of programming that is likely to be found easy by most users regardless of whether or not they have done programming before.
- 4) *Improved scalability.* It is possible to install a MoSeS service (such as the analysis service) on multiple machines, and then perform load-balancing by dynamically binding to available/free service.
- 5) *Fault-tolerance.* Related to the benefits of improved scalability mentioned above, a level of fault-tolerance can be introduced into processing by, for example, invoking multiple MoSeS services in parallel and either cross-checking their results or else using the results of the first service to fail without raising an exception. This can also lead to performance gains, as multiple services can be invoked and the first result to be returned can be fed directly into an ongoing workflow.
- 6) *Ease of maintenance.* In spite of the increased number of entry points into the system, maintenance of core MoSeS functionality should not be affected, as changes made to each service will be reflected in each interface due to a clear separation between presentation and application logic layers of the system

A drawback to service-enabling MoSeS is the potentially slower performance of individual tasks (note this is unrelated to the scalability issue), mainly when passing large volumes of data using the SOAP (SOAP Version 1.2, 2007) messaging protocol. Typically, data passed between different MoSeS services would need to be serialised into XML form, sent over HTTP, and then deserialised at the other end of the connection, resulting in noticeable delays when considering the size of data that is often required to be passed between services. A partial solution to this problem has been developed, whereby each MoSeS service stores data for a particular user session in SRB space, and invokes other MoSeS services by sending a reference to this data, which can then be transmitted in binary form.

The service-oriented architecture for MoSeS is shown in Figure 1. As can be seen, MoSeS web services can be invoked by any number of end user interfaces, including Java GUIs, JSR-168 portlets, workflow enactment engines, etc. Services can be distributed remotely as all data interactions are performed either through direct access with distributed resources over HTTP, or else through accessing the MoSeS SRB cluster, which in turn stores the results of the MoSeS demographic and forecasting modules.

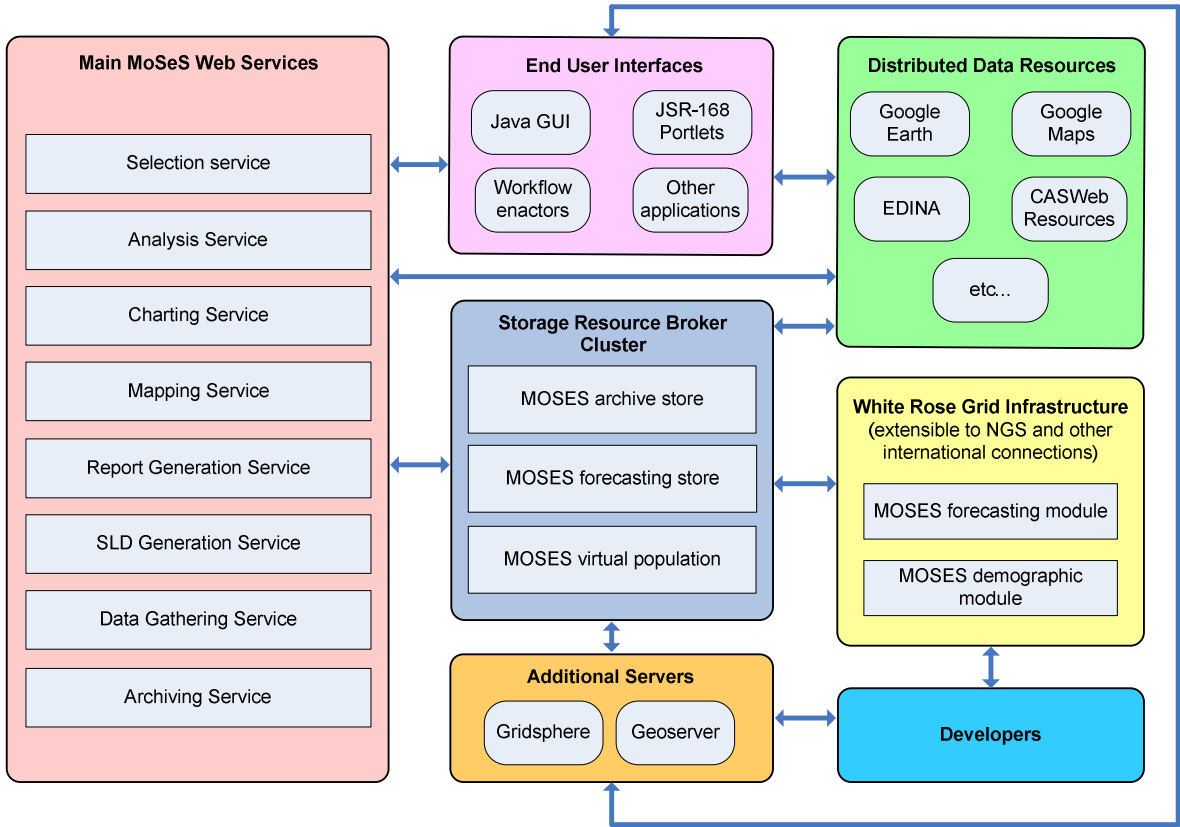


Figure 1. The service-oriented MoSeS architecture.

Due to the new ability to use multiple end user interfaces within MoSeS, this architecture has also been designed to exploit a number of Web 2.0 technologies, particularly those related to mapping, in order to provide a more dynamic and rich user experience. Web 2.0 is defined by O’Reilly as “the business revolution in the computer industry caused by the move to the internet as platform, and an attempt to understand the rules for success on that new platform” (O’Reilly, 2006). The use of these technologies within MoSeS is discussed in section 5.

3. Simulations

In order to illustrate the event-driven simulation model used by MoSeS, the urban area of Leeds - a city of 730,000 people in the north of England – has been used. The Leeds area is used for illustrative purposes, but is completely generalisable between local areas across the country.

The base year simulation is established by reweighting the Household Sample of Anonymised Records (HSAR) to individual census wards in Leeds Metropolitan District. HSAR comprises a 1% sample of households from the UK Census of 2001 in which the census questionnaires are completely enumerated for households and their constituent individuals. The essential mechanism for protection of the anonymity of individuals is through reduction in the geographical resolution of data - thus each household can only be identified to the level of a Standard Region (i.e. South-West, Yorkshire and Humberside etc), which limits the value of the source data for the purpose of spatial microsimulation.

Our approach is then to synthesise records from the HSAR in accordance with the structure of individual wards. Each ward population is therefore a unique extract or ‘re-sampling’ from the HSAR in accordance with the local geography. Households are sampled ‘without replacement’ from the parent distribution and are therefore unique within an area, although it is not only possible but necessary that some households will be duplicated in the Leeds area, as Leeds accounts for more than 1% of the UK population, and so there are more households in the city than in the household sample.

The reweighting procedure is based on successive proportional sampling from the HSAR. Initially, the desirability of selecting a candidate from the HSAR is random. From this random selection, the composition of the sample is compared to key population distributions from the 2001 Census Small Area Statistics (SAS). The weights are adjusted to increase the likelihood of selection for population sub-groups which are under-represented, and vice versa. This adjustment process continues until the observed and predicted population distributions are similar within a specified tolerance. The simulation model is anchored in a base year population for the year 2001. This is necessary in order to exploit the richness of the UK census of population and households for the purpose of the simulation; there are no appropriate data sources to facilitate the preparation of a more up-to-date base population.

The objective of dynamic modelling is simply to project the population forwards in time. For such purposes, the most commonly used approaches have been cohort-based ‘macrosimulation’, in which the population is divided into categories, and multipliers – such as mortality rates or birth rates - are applied to those individual categories. However these methods are problematic whenever a relatively rich set of population attributes is involved, as the number of categories begins to grow exponentially. Van Imhoff and Post (1998) present an example in which the population of France is represented more efficiently through an individual microsimulation model even though age, marital status and place of residence are the only variables.

A number of projects have therefore attempted to build dynamic microsimulation models, particularly for economic applications (e.g. CORSIM, SVERIGE) but also for social and anthropological applications to problems of kinship and community (Murphy - 2004, van Imhoff and Post - 1998). However the only examples of demographic projection with spatial microsimulation use the technique of ‘static ageing’, in which a base population (the HSAR in our previous example) is resampled in the context of independent estimates of future population change (e.g. Ballas et al, 2004). Therefore this method provides no means to monitor the dynamics of change within a population, and ignores the benefits of dynamic microsimulation as a means for the projection itself. Some authors have seen fit to make a distinction between

policy and pedagogic applications of microsimulation (van Imhoff and Post, 1998). In this context, it is argued that the accuracy of models with a policy orientation need to be validated in a real world context, whereas pedagogic models need only to reproduce local interactions within the population. We remain sceptical of such a distinction, and view the process of validation as essential if robust conclusions are to be drawn which are independent of the model as artifice. The MoSeS dynamic model consists of seven components:

- *Mortality.* The mortality component of the dynamic model predicts the expected number of deaths within each UK census ward, using data obtained from census data and ONS Vital Statistics. At each time period, a survival probability is applied to each individual on the basis of age, gender and location. The model is run in annual time increments, and therefore the ageing rule for all survivors is that they become a single year older in each time step.
- *Fertility.* Ward-specific fertility rates are derived in a rather similar way to the mortality rates, with national rates again being localised in accordance with ONS Vital Statistics.
- *Health Status.* Individual health states are recorded in the HSAR and SAS within five categories ranging from very poor to very good, but for both convenience and robust estimation these are reduced to three categories of 'poor', 'medium' and 'good' within the simulation. For each individual, the probability of a change in health status is assumed to be dependent on current health status, age and gender. These rates of change are derived from the BHPS.
- *Household Formation.* Changes in household composition are determined by four processes in the model. These are the formation of new unions (including marriage), the dissolution of existing unions, movements in which one or more persons leaves a household, and the death of a household member.
- *Migration.* In order to accommodate migration, the model has two important features: a stock of houses which are independent from the households which occupy them, and a location search process which is mediated through an aggregate spatial interaction model (SIM) of migration. This model recognises that different households – according to their size, composition and age – have different housing preferences and search horizons.
- *Module Sequence.* In this application, we have adopted a fixed order of events in the sequence fertility, health change, mortality, migration, household formation. Fertility and health change both appear before mortality, since there are still risks of infant mortality, and although death rates in our model are independent of health status there is a logical connection between mortality and deteriorating health. Since many new households are formed in association with the migration process, it makes sense to consolidate household structures once a move has taken place, while recognising that a more desirable option would be to consider these two processes simultaneously
- *Student Migration.* In wards where student migration has a great impact, the microsimulation failed to reproduce the student population renewal and they grow old in the areas as other people do. We know however for a fact that students tend to only stay in such areas during the period of their study and then leave while other new students move in. Due to the replenishment of student population each year, the population in such wards stays younger than that in other wards. A hybrid approach combining ABM techniques is therefore adopted to strengthen the modeling of such subtlety of the local migration patterns and the behavior modeling of the student migrants.

4. Latest Simulation Results

Some key results from the current implementation of the Moses dynamic spatial microsimulation model are described in this section. Results are presented over a 25 year projection horizon from 2004 to 2029. In relation to our previous discussion about assumptions, survival rates are expected to improve by 1% in each age-gender-location combination in each year of the simulation. Fertility rates are expected to increase by five percentage points for each demographic sub-group in each year to 2011 before stabilizing.

Firstly, we note some basic demographic trends. Regarding the age composition of the population, we find a surge of 20% or more in each of the elderly cohorts 65-74, 75-84 and 85+. The school age cohorts all end up reasonably close to where they started, albeit by rather different trajectories. Current fertility levels in Leeds as elsewhere are close to their historical low, but beginning to rebound. An assumption of increased fertility levels is offset by lower numbers moving into the major child-bearing age groups. These results are very much in line with the Office for National Statistics expectations for the Leeds area (ONS, 2007).

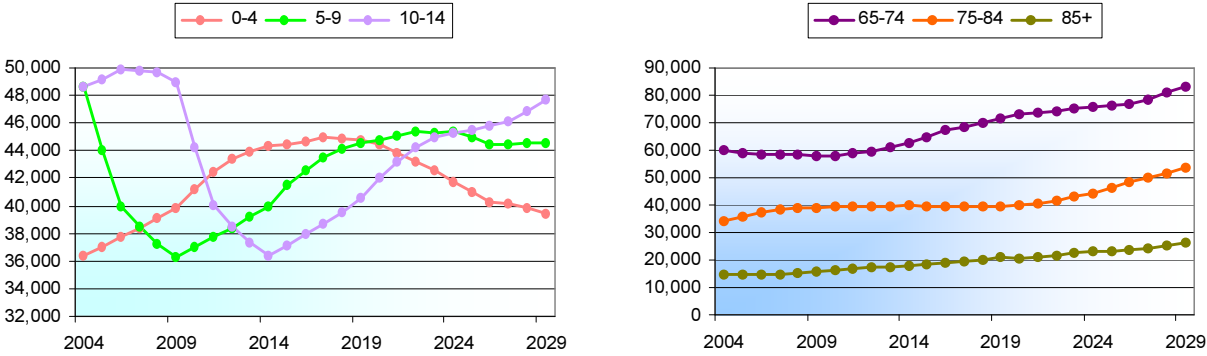


Figure 2. Projections by age cohort: a) School age; b) Elderly

Ethnic group projections show substantial growth in all minority groups. These are the product of two effects – ongoing net migration of minorities into Leeds and a demographic bulge in the younger, more fecund age groups. Note that fertility rates are uniform between ethnic groups with the same age and marital status profiles. These projections are again in line with estimates produced on behalf of the local development agency (Rees, Stillwell, Boden, 2007).

Table 1. Ethnic Minority Projections

	2006	2011	2016	2021	2026	2031
UK	671334	675994	687030	697905	702397	702929
New Commonwealth - Africa and Caribbean	13464	14321	15815	17786	19489	20802
New Commonwealth - Asia	32515	34987	39457	44709	49551	54118
Others	21245	24616	29162	34236	39900	46556

Spatial disaggregation in the growth of elderly populations is illustrated in figure 3. Here the notable trends are an accumulation in areas where the older age groups are most heavily concentrated already, notably the Wharfe valley across the northern edge of the city, Cookridge and Halton wards. However there is significant growth in all areas through ageing in situ: the migration process does not simply relocate all the pensionable age groups into a small number of retirement areas. More detailed effects can now be explored through secondary modeling. In the next illustration, we have used a data linkage model (Zuo, 2007) in order to match records from the BHPS with individual HSAR records, and from this we have extracted disability rates. The total individuals with a disability is then summed back to census wards and plotted as

Figure 4. Here we can see some relationship between disability and old age in the wards of North and Cookridge, but the dominant pattern is one of centralization. The poorer socio-economic groups, often associated with low economic activity rates and high unemployment, in the central areas clearly experience much higher like-for-like disability rates than their more affluent suburban counterparts.

Rates of disability are projected into the future in two ways. First, the projected individuals from the dynamic microsimulation are again matched with individuals from the BHPS 2004 to provide an estimate of disability with current health patterns. In this scenario, the proportion of disabled people in Leeds rises from 9.1% to 14.1% in a 25 year period. In the second estimate, we assume that for populations over 40 individual health improves steadily so that in 25 years time everyone is ‘five health years’ younger than at present – for example, a typical person aged 65 in 2031 has the average health characteristics of a 60 year old in 2006.

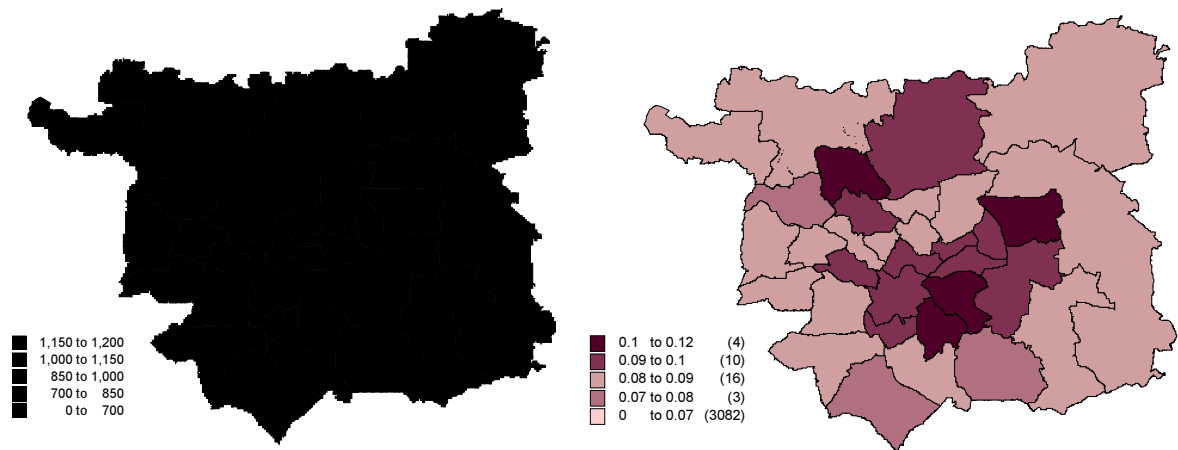


Figure 3. Growth in population 85+ to 2031. Figure 4. Disability in Leeds, 2006

Given current concerns over conditions such as obesity and diabetes, such an assumption may well be optimistic, but even in this scenario the disabled population of Leeds grows from 51,600 to 70,400 – a 36% increase. On this metric it is clear that social services in Leeds will be under acute pressure over the next 25 years, and that the spatial consequences of increasing need will be uneven.

5. Web 2.0 Visualisations

MoSeS has been given the capability to render simulation results using Web 2.0 technology, allowing for the use of both dynamic 2D and 3D maps of results. This is performed primarily through the use of *GeoServer* technology (<http://geoserver.org/display/GEOS/Welcome>). Geoserver is an open source, freely available implementation of the *Open Geospatial Consortium* (OGC) Web Feature Service (WFS) (<http://www.opengeospatial.org/standards/wfs>) and Web Map Service (WMS) (<http://www.opengeospatial.org/standards/wms>) specifications, and allows map data to be uploaded and served out in a variety of formats.

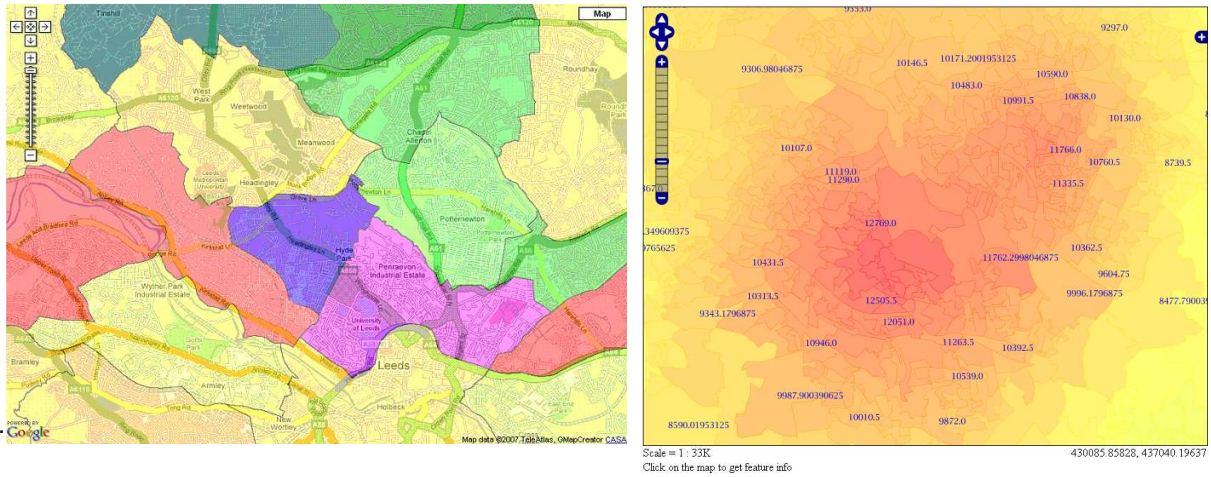


Figure 5. The MoSeS Google Maps and OpenLayers interfaces.

The MoSeS Google Maps and OpenLayers interfaces are shown in Figure 4, with the Google Maps interface on the left and the OpenLayers interface on the right. In addition to the OpenLayers interface, another highly useful format served out by Geoserver is *Keyhole Markup Language (KML)* [14], an XML-based language schema for expressing geographic annotation and visualization on two-dimensional maps and three-dimensional Earth browsers. KML is the native file format of the Google Earth [15] application, and when the KML of a MoSeS map is combined with a generated SLD, impressive 3D visualisations of MoSeS data are possible, as shown in Figure 5. Once Google Earth is loaded, additional information, such as place names, roads, etc, can be layered on top of this visualization by the end user, using the Google Earth application interface.

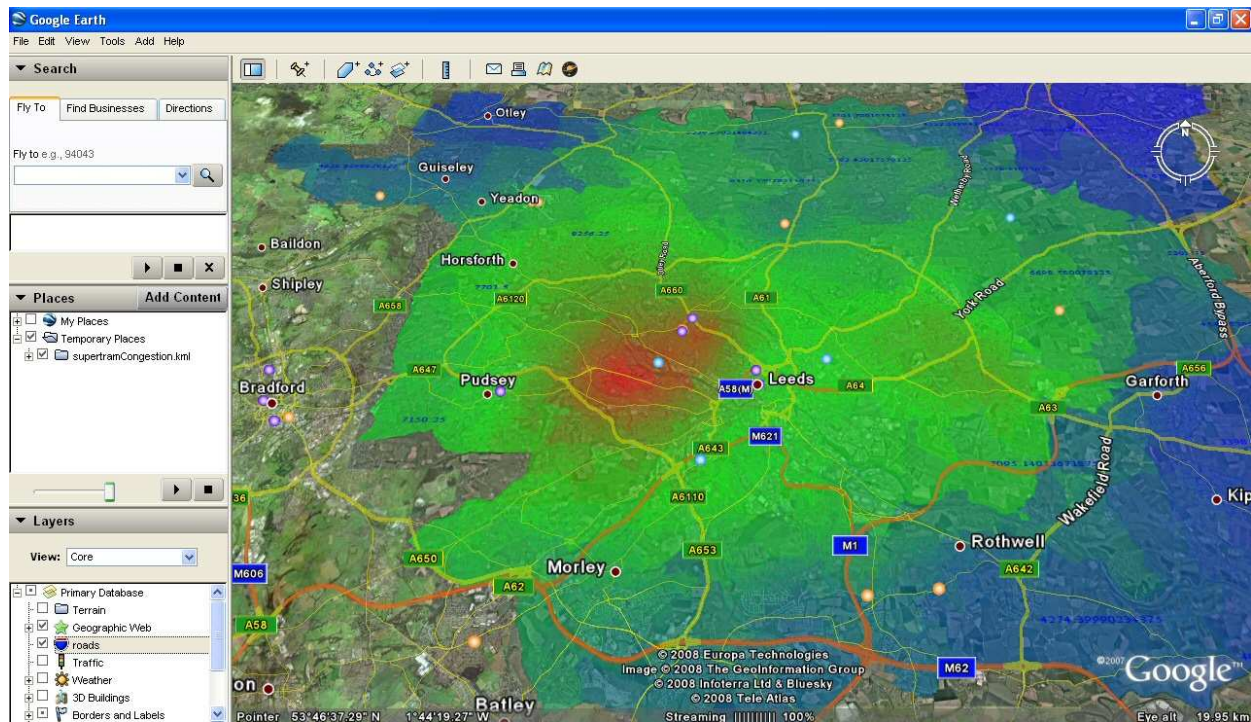


Figure 6. MoSeS data visualised in Google Earth.
Source: 2001 Census, Output Area Boundaries. Crown copyright 2003

6. Conclusions

This paper has described the architecture, simulation methodology, and latest results obtained by the MoSeS (Modelling and Simulation for e-Social Science) project at the University of Leeds, which uses e-science technologies to improve existing social science tools by greatly ease the integration of new datasets, and quickly take advantage of new and large-scale computing resources in a dynamic fashion. The functionality underpinning MoSeS is service-enabled, giving a further advantage over existing tools by allowing for multiple entry points to the system, richer user interfaces, increased user flexibility, improved scalability, fault-tolerance, and easier system maintenance. A synthetic representation of the entire Leeds population has been generated from publicly available datasets. Using an events driven model which simulates discrete demographic processes, the population has been projected 25 years into the future. Whilst the approach is grounded in the methods of microsimulation, concepts from spatial interaction modeling and agent-based systems are incorporated in an innovative way. Although appropriate simplifying assumptions have been introduced, the model still incorporates a great many parameters and assumptions. Web 2.0 technologies have been employed to allow for a more dynamic and rich user experience when visualising MoSeS results; this has been achieved through the use of third party server software such as Geoserver, in addition to a bespoke MoSeS service for the generation of mapping styles, and allows MoSeS data to be visualised in Google Maps, Google Earth, and with OpenLayers.

7. Acknowledgements

MoSeS is funded by ESRC Grant RES-149-25-0034.

8. References

- Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., and Rossiter, D. (2004) 'SimBritain: A SpatialMicrosimulation Approach to Population Dynamics', *Population, Space and Place*, 11, 13-34.
- Van Imhoff, E. & Post, W. 1998. 'Microsimulation methods for population projection.', *Population: An English Selection*, 10: 97-138.
- Murphy, M (2004): 'Tracing very long-term kinship networks using SOCSIM', *Demographic Research*, 10, 171-196.
- Office for National Statistics (2007) *Sub-national population projections, 2004-2029*, HMSO, London.
- O'Reilly, T. (2006): *Web 2.0 Compact Definition: Trying Again*, <http://radar.oreilly.com/archives/2006/12/web-20-compact-definition-tryi.html>
- Rees, P., Stillwell, J. and Boden, P. (2007) *Ethnic projections for West Yorkshire*, School of Geography, University of Leeds for Yorkshire Forward.
- W3C Recommendation (2007): *SOAP Version 1.2 Specification*, <http://www.w3.org/TR/soap12-part1/>
- Zuo, C. (2007) *A model of house prices in the Leeds area*, unpublished MSc thesis, School of Geography, University of Leeds.

9. Figure Captions

Figure 1: The service-oriented MoSeS architecture

Figure 2: Projections by age cohort: a) School age; b) Elderly

Figure 3: Growth in population 85+ to 2031

Figure 4: Disability in Leeds, 2006

Figure 5: The MoSeS Google Maps and OpenLayers interfaces.

Figure 6: MoSeS data visualised in Google Earth