eIUS: Demographic Modelling Experience Report

In the text: Some barriers in '{{…}}' to provide more contextual information.

Interviewee profiles (group interview; two interviewees)

Both are researcher in the MoSeS (Modelling and Simulation for e-Social Science) project (NCeSS node) and work at the School of Geography at the University of Leeds;

Time spent in research

Researcher one: 50% on research and 50% on teaching

Researcher two: nearly full time on research

Research area

MoSeS – in the context of e-Science and e-Social Science and from the perspective of Geography, computational Geography (modelling and simulation of demographic developments), spatial analysis, urban planning and policy – is the biggest project they are involved in together with other activities/projects in "the same general area, reasonably focused but not directly associated with e-science, e-social science".

Research question(s)

"The high level branding that we sometimes use is the SimCity for real type thing (..), we are building simulations of cities that can potentially help planners, policy makers to make decisions, find decisions about cities and see what the impacts of those might be, so I mean academically it's a kind of the interface between geography and urban planning, I would say, a kind of site urban planning if you like".

"(..) in essence the research question is still can we do it? You know we want to demonstrate that it's possible to build realistic simulations, and you know, that actually can tell you something, is worth knowing about, you know the future or the impact of policy changes (..)".

"There is lots of simulation side of things, there is also something we call Special Decision Support Systems, so it's how computational technologies can support decision making (..). I think, in computer science, I mean in number of areas, we are particularly interested in geographical context so that involves (..) data integration; (..) you are bringing together various kinds of data sources in order to be able to do this sort of work [which] potentially involves (..) to be able to visualize what the impacts of some of these things are and it potentially involves quite (..) intensive simulation, (..) a lot of number crunching". But social science and methodological questions have to be addressed as well to apply those models and simulations.

The end users are "kind of policy makers, people in the real world" – "we are not really to develop something for academic research" in itself.

Actual geographical questions: “I am a geographer although my training in geography is in mathematical modelling or geographical modelling; (..) we approach it [i.e. the research] from a (..) methodological point of view, what we are interested in is how you do it and whether it can be done (..); one of my colleagues in a project is a demographer, so he is more into substantively (..) what are the main populations trends and processes and migrations patterns (..), cities and all that side of things.” (see also first paragraph under ‘Collaboration’)

Research Lifecycle

Start of the research process (Literature Review), already includes data collection required for development

The start of the research process is driven by the needs (e.g. the underlying data) for the development of the tools and models for demographic modelling: “That’s the first thing, so that’s we are trying to find the data that we are going to use, but in the proposal for MoSeS it was already stipulated that we’d use 2001 census aggregate statistics and samples on those records as they become available.”

Questions and answer in the context of effort needed in addressing new research question in the research process:

Interview QUESTION: ‘The service generation step (acquiring the data and creating the model for running the simulation to come up with policy evidence implications) is the step you will have to do for each new substantive research question? So each researcher’s dataset that would come in, each (..) new version of census or panel, each different slice through variables of interest and so on?’

“Not necessarily, (..) that’s one of the things in effect that we are looking at, I mean you have got two options, there is one you can create, you can do a general version, you can do it once or you can do kind of customize relations, but the problem is there is quite a lot of information that we are dealing with here, so you effectively (..) [try] to optimize lots of different distributions simultaneously, so what Andy is saying, effectively is that if you do a kind, like a general version, you know for health and transport and housing, then you are not necessarily going to get a quite a definition in terms of (..) everything is in a right place as if you did those things individually, so you might want to do it, you might want to re-initialise the whole thing; now this is a different kind of aspect to (..) the other thing you can do (.), you can generalise, if you like, sort of part of the data and so (..) we did quite a lot with this sort of data linkage, they say in that case you create your initial population and they say right now come along with some health service record or something, and then actually merge those sets of records in some way and (..) that’s a so much more efficient process and so again in that case you wouldn’t be needing to regenerate initial population”.

Data collection process into modelling

Datasets come from Mimas, Edina, ESDS as well from the Office for National Statistics (i.e. their ONS provider), e.g. regarding Samples of Anonymised Records data (household SAR) or general household survey data (GHS). The data mostly comes as Comma Separated Values (CSV) files.

Interactive services to browse/search the data online: CASWEB (http://casweb.mimas.ac.uk/); “The census aggregate statistics and (..) metadata (.)

which explains (.) all the various variables and it is tied up in an interface there, and so we can copy and paste from that interface as we are using, browsing the data and then we get the data files which just have a sort of arbitrary code for each of the variables which you have to reference back to this metadata to know what it is you actually mean." In general it is possible to download the raw data or to copy and paste from the browsing data – and in the end the interface is not important: "No, the data is, I don't care how we get it."

{{BUT: "They add value to it but it's tied up into its interface at the moment (..), there is no automatic loader that you can write for a table, you have to sort of learn for each table.}}

Data from previous simulations is fed into new models as well: "Yes, that's done in principle, you know that's part of the technology archive that (.) [a **colleague** in Leeds] is working on at the moment (..). It's pretty much all home made".

A student form China will **join the team** in due course and work on a sub-project on additional applications for the transport domain: "I envisage that he may be using some colour package transport someway to some degree (..). (..) there could be an element of, (..) embedding further application software within what we do, but most of our stuff (..) is home made."

Java is used as the language of choice in MoSeS, but a **colleague** (also in Leeds, but working in a different department) with a computer science background also uses C++: "He needed something that got to work quickly, it's a small function then and it's too slow at the moment, then he has talked about implementing that in (..) a function that works faster and we can glue together with the Java".

Data analysis/Simulation

Statistical analysis: SPSS is sometimes used to look into the data, to pre-process it, but "that's kind of external to the actual MoSeS application". **Data**: "That would be both the general household survey and the British household panel survey".

Database: **Another project member** is mostly working on the database technology (Apache Derby: http://db.apache.org/derby/) from Ireland, i.e. "building the dynamic model of the demographic part of this work". In the database "there are individual records, (..) [a] rich list of individuals, if you like, characterised by the kinds of housing that they are live in and by the health status and the jobs they are doing and etc, etc.". This is the underlying, anonymised data used for the simulation.

"But in effect it's a national population of these people that we are interested in, so we are trying to simulate transitions, so how people (..) [have] to move house and how do we (..) [address the] litigation system, because all that kind of stuff (.) is quite complex, (..) quite complicated in terms of the actual simulations itself (..)."

Simulation time: E.g. for a "toy model at the moment for Leeds and so that's about 1% of the country, and we've not managed to sort of parallelise it, so just working on the sort of basic, average PC of today" it will take a couple of weeks.

{{"Difficult to parallelize, some parts of it, especially the migration component, because you are trying to work out how many people are moving from one region to go to another, but obviously that contrary to how many people can move in to that region, so it's difficult."}}

Use of Grid HPC computing (used for the population generation before the actual simulation can be started): "We use NGS resources and WhiteRose grid resources to do the population and initialisation which is creating the SAR population for 2001 and that again was a computationally intensive and data intensive task and (..) we've worked with an expert who used to be in Manchester and he's written some stage and stuff that allows us to use MPJ express [Java for HPC: http://mpj-express.org/] on those resources." Grids ("we can use whole 128 nodes on the core side (..), it's more than 200000 out per areas for the UK and so we take 128 of those and run something that takes 5 minutes for each one we start getting results back in") are used for the population generation/recreation which has to be done before the actual simulation can be started (again), as this is a parallel-computing activity which benefits from HPC: "NGS, we sort of done a calculation, we've used 128 nodes on each core of 4 core sites then we can get the results out in **two weeks**." – The operation would take about **three months** on the local cluster in the department with 28 nodes!

More on population generation: "Yes, that needs to be done first and you can do that in many different ways with many different random sets and different things that you are trying to constrain or optimise onto; so for certain applications, we might be interested more in getting our health variables correct within the initialised dataset and so to do what we would like to, we'd try to optimise constraints onto specific census variables to get those things more right and everything else, because you cannot get everything exactly right, and so this initialisation thing, there is two things we do with constraints, we control constraints (..) for each area so we have to get these things exactly right, there is only a certain number of variables, the more variables you include the less chance you have of being able to control constraints from the sets that are available from the SAR, and then after you've done, (..) the other things are optimisation constraints – and these are the things you try to get right."

Evaluating simulation results (are the results accurate): "Essentially there are two ways of validation (.) based on simulation (..) one is against real data that you know about but you haven't kind of factored it in to the simulation process and the other is, (..) comparing different techniques, so if you are basically trying to do the same sort of things, we are trying and optimise (..) against people with those health characteristics, by three months methods as opposed to the five minutes method (..)." Comparing different techniques means tracking variables across city models, applying "stratified type analysis, regression analysis, statistical techniques you can use to (..) compute the areas between various sets of the distributions".

Tools used to collate data, analyse and visualise it are mainly Excel and SPSS complemented by "a bit of stuff, that (..) I write myself". Graphs are also created automatically as J3 (Java) charts.

Collaboration

About a colleague working on the classical geographic demographical side of things, not being involved in or using computation (see also last paragraph under 'Research question(s)'): "He (.) actually effectively is advising us on development of population, (..) certain technical companions of the non-development, I mean that is also a sort of technical process. Say, he has got much more substantive interest in the actual planning."

A collaborator and future user in the project is an expert in healthcare economics who is interested in applications in that area.

Transport studies are another area in which they are collaborating with an expert developing models for this domain.

F2f meetings and discussions take place regularly between the two interviewees at Leeds, sometimes joined by another collaborator (domain expert) based there.

"(..) there is a group at Sheffield one of our actual research student is there and actually I have a joint PhD student (..) scheme with Sheffield, now he is doing some sort of related work but I mean we don't directly interact on this stuff, (..) primarily because we are not funded to do so."

Already mentioned before (under 'Data collection process into modelling'): One colleague works in Ireland most of the time while another one works in another department (Computer Science) at Leeds University; a student form China will join the team in due course and work on a sub-project.

In the near future (at the time of the interview) MoSeS will be joined up with another NCeSS node, GeoVUE (based in London), to become the GENeSIS node in a new funding phase.

Use of tools in collaboration:

"Anything that I choose to mark in a meeting, it's usually (..) [in] a digital document, although we might print things out, and every time I have a meeting, then that information goes onto the webpage (..)."

One of the interviewees extensively uses wikis and blogs in his everyday research: "Wikis are very useful, a basic way of collaborating and compiling information (..)". He logs his daily activities as well as project meetings (minutes) and there like on his web page (web log; blog log) in html and this way keeps a kind of history of his work and the project. Other project members and collaborators quite often read this documentation; also "I am trying to encourage them all to contribute to that activity". The interviewee started using the NCeSS Sakai portal, but at the moment it is not used by others in the project.

Dissemination

Publications: Work has been published "in conference proceedings, we've got a couple of book chapters, there is something in a journal called 'Computers, Environment and Urban Systems (CEUS)'."

"The computer science group [in the project], they tend to go through the (..) conferences, (..) those proceedings tend to get published, other than that, I mean we are not at the phase of (..) writing up, as I said, (.) it's (.) conference presentations and proceedings that we mostly publish."

Software: Software written in the project is published on the school of geography web page under LGPL with links on the individual or project web pages – which also in cases would be referenced in papers.